

Linguistic Typology at the Crossroads



ISSN 2785-0943

Volume 6 – issue 1 – 2026

Issue DOI: <https://doi.org/10.60923/issn.2785-0943/v6-n1-2026>

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

Caterina Mauri (University of Bologna, Editor in chief)

Andrea Sansò (University of Salerno, Editor in chief)

Francesca Di Garbo (University of Aix-Marseille)

Nicola Grandi (University of Bologna)

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)

Production editors

Valentina Di Falco (Independent editor)
Eleonora Zucchini (Masaryk University)

Scientific Board

Giorgio Francesco Arcodia (Università Ca' Foscari, Venice)
Peter Arkadiev (Johannes-Gutenberg University of Mainz)
Gilles Authier (École Pratique des Hautes Études, Paris)
Silvia Ballarè (University of Bologna)
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)

Assistant production editors

Elia Calligari (University of Pavia)
Antonia Russo (University of Bergamo)
Filippo Sergio (University of Bologna)
Silvia Zampetta (University of Pavia)

Responsible Editor

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



Linguistic Typology

at the Crossroads



CONTENTS

Affirmative markers – A cross-linguistic study of overspecification

Tim Zingler ----- 1–54

Habitual markers and their lexical sources across languages

Sune Gregersen, Eva van Lier----- 55–100

From construction to suffix: Grammaticalization and areal diffusion of jussives and optatives based on ‘say’ in Daghestan

Nina Dobrushina ----- 101–140

Basic verbs in European languages

Åke Viberg----- 141–214

Manifesto Articles

Linguistic typology and the cognitive science of non-WEIRD societies: The role of memory types

Antonio Benítez-Burraco, Sihan Chen, David Gil----- 215–235

Errata

Corrigendum. Is complementation a universal strategy? A cross-linguistic corpus study

Nicholas Evans, Danielle Barth, Wayan Arka, Henrik Bergqvist, Christian Döhler, Sonja Gipper, Yukinori Kimoto, Dominique Knuchel, Daniel Majchrzak, Hitomi Ōno, Eka Pratiwi, Saskia Van Putten, Andrea C. Schalley, Asako Shiohara, Stefan Schnell, Yanti----- 236–237



Affirmative markers: A cross-linguistic study of overspecification

TIM ZINGLER

UNIVERSITY OF INNSBRUCK

Submitted: 16/05/2025 Revised version: 18/09/2025

Accepted: 07/10/2025 Published: 24/7/2026



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

Abstract

Based on a worldwide sample of 604 languages from 179 families, this study shows that affirmative markers, the functional opposite of standard negators, are highly infrequent and possibly ‘rara.’ This confirms previous impressionistic claims. Nine languages with one affirmative marker each were found. Their geographical distribution (a major cluster in eastern Africa), exponence (five of the nine markers are prefixes), and possible diachronic emergence are discussed. The existence of affirmative markers is explained against the background of ‘overspecification,’ operationalized here as the combination of a high degree of verbal synthesis and a low number of zero markers throughout the system. Languages with affirmative markers generally have a large number of morphological categories/slots and also show a tendency to overtly mark other meanings often zero-marked cross-linguistically (e.g., nominative, present tense, third person). The results suggest that some languages prefer explicitness to economy, which informs different aspects of typology.

Keywords: Morphology; morphosyntax; rarum; polarity; affixation; negation.

1. Introduction

One of the primary motivations underlying linguistic typology is the recognition that languages express certain meanings much more commonly than others as part of their grammatical systems. The traditional approach within which these asymmetries were

framed was the concept of markedness, which, among other things, assumes a correlation between semantic load and formal coding. Roughly speaking, the most basic value of a grammatical category is expected to receive the least amount of coding. That value is thus ‘unmarked,’ while the other members of the paradigm are ‘marked.’ However, the empirical facts underlying markedness have always been on a firmer footing than the actual theoretical account. For example, Greenberg (2005 [1966]: 65–69), in arguably the definitive work on typological markedness, already pointed out that what is conceived of as the unmarked value within a category is typically just the most token-frequent one. This account has since been elaborated upon by, for instance, Fenk-Oczlon (1991) and Haspelmath (2006, 2021). These authors argue that, as far as explanations go, token frequency is superior to markedness and the often vague semantic assumptions that it is based on. This is because the link between token frequency and formal shortness was already clearly established by Zipf (1965 [1935]). To the extent that this line of argument holds, then, markedness is best reduced to a descriptive term for linguistic forms. That is, a nominal form with an overt plural morph is marked, whereas the corresponding singular form without an overt singular morph is unmarked, regardless of whether and why the notion of singularity might be more cognitively basic than the notion of plurality. Unless stated otherwise, the present work will use the terms ‘marked’ and ‘unmarked/zero-marked’ in this formal sense. That is, I will not offer any claims about the potential psychological underpinnings of certain semantic values, and thus I will not try to provide an argument for markedness and/or against frequency.

Of all the marking asymmetries within grammatical categories, none might be more salient than the one found in the domain of polarity, i.e., the contrast between negative and affirmative (or positive). Dixon (2012: 90) makes the strongest possible case by claiming that “[n]egation is a term in a two-member polarity system {positive, negative}. ‘Positive’ is always the formally unmarked term, with zero realization.”¹ If true, this would mean that no language has overt affirmative markers. This would be remarkable since absolute universals are known to be rare, and fieldwork on previously undocumented languages continues to reveal structures thought not to exist (and occasionally fails to find structures thought to exist universally). On the other hand, the absence of affirmative markers from the languages of the world would

¹ An anonymous reviewer points out that this conception of polarity, which does not consider interrogative or conditional constructions (etc.), would be considered too narrow nowadays. Since it is sufficient for my argument, however, I will adopt it for the purposes of this paper.

be understandable in that the binary structure of polarity (Miestamo 2005: 2, 39) paves the way for the efficient covert coding of one of the two values of the dichotomy.

Dryer (2005: 454) points out more explicitly than Dixon (2012: 90) that all languages have negative morphs (not including intonation, word order, etc.).² This claim is essentially also made by Bybee (2006: 180), who lists the availability of a negative construction as one of the few absolute universals that have been securely established; cf. also Dahl (1979: 79), Zeijlstra (2007: 499) and Shcherbakova et al. (2023: 2) for similar statements. Greenberg's (2005 [1966]: 50) view is more nuanced. He posits that negators are universally available but points out that there are also some languages with dedicated affirmative markers.³ This description has been echoed by Haspelmath (2021: 614). With respect to the affirmative function, Lehmann (2004: 1845) suggests that it is “normally unmarked,” whereas Wurzel (2001: 184) states that it would be inconceivable to have a language with a zero-marked negator but an overt affirmative marker (cf. also Givón 1991: 350; Bybee 2011: 140). This assumption hints at an implicational hierarchy such that each language that gives overt expression to the unmarked value as traditionally understood (affirmative) will also do so with respect to the marked one (negative).

Based on the above, there appears to be a broad consensus concerning the structure of polarity systems, with the only difference coming down to whether affirmative markers are very rare or non-existent. In this context, it should be mentioned that the category of polarity seems to confirm the predictions of the frequency-based approach to coding asymmetries. That is, negated clauses are much less token-frequent than affirmative ones (cf. Givón 2001: 43, 373 for English data). This is presumably because communication is more efficient if it focuses on the few states and events that do, did, or will hold rather than on the countless states and events that do, did, or will not (cf. Hopper & Thompson 1980: 287).

Yet, while the notion of token frequency may serve as an explanation in typological projects, the empirical groundwork will have to establish the type frequencies of the

² This is not contradicted by certain zero negator constructions in some Dravidian languages since the relevant languages have overt negators in other constructions (Miestamo 2010; Pilot-Raichoor 2010).

³ He mentions Vietnamese (vie; Austroasiatic, Vietic) as one of the languages with an overt affirmative marker, but I found no such element in any of the three Vietnamese grammars that I consulted (Thompson 1987; Nguyễn 1997; Ngo 2021). Cf. Section 2 for my definition of affirmative markers and Section 3 for the database.

pattern(s) at issue. As such, it is noteworthy that this question seems not to have been addressed for affirmative markers before. For instance, if they are indeed rare, one would expect them to be discussed in Frans Plank's *Grammatisches Raritätenkabinett*.⁴ However, the only entry that includes the term 'affirmative' in that database is for "negation expressed negatively, by omission of material present in affirmative clause (marked realis vs. unmarked irrealis mood inflection)." This is not the sense intended here, where an affirmative marker expones a value of the category of polarity and is an overtly coded means that expresses the opposite of standard negation (cf. Section 2 for details). This disregard for affirmative markers could mean different things. On the one hand, they might be too common to count as a rarity, pace the intuition of most linguists. On the other hand, the complete lack of affirmative markers might be taken for granted, following Dixon (2012: 90) cited above. The latter idea finds anecdotal support in that the relevant chapters in reference grammars and textbooks are typically called 'negation' rather than 'polarity,' which suggests that negators are the only overt expressions expected in polarity systems.

The present work takes the lack of previous research on the worldwide distribution of affirmative markers as its starting point and conducts a cross-linguistic study based on a large sample of reference grammars. That is, it will investigate if, where, and how commonly affirmative markers are (claimed to be) found, which will help to clarify whether the absolute universal posited by Dixon (2012: 90) is valid. If that claim turns out to be false, affirmative markers might instead have to be classified as 'rara,' which have also drawn attention in the typological literature (cf. Section 2).

However, the interest pursued here goes beyond the empirical dimension in that this study will also investigate a specific hypothesis. As with the typological distribution of any phenomenon, the question arises as to why languages show it at all. This, of course, is difficult to answer in any case, and it might be argued that an explanation would be even harder to come by in the case of a (presumably) rare structure. Yet, there is a strand of morphological typology that happens to be instructive in the present context. McWhorter (2005) discusses how the structure of languages whose cultural transmission was never interrupted differs from that of creoles, which on his account do derive from scenarios of interrupted transmission.

⁴ I have not found an official way to cite this resource. It can be accessed at <https://typo.uni-konstanz.de/raraalt/intro/> (last access: March 2, 2026).

In the process, he develops the concept of ‘overspecification,’ which he defines as follows:⁵

Any given language gives overt manifestation to particular underlying semantic or syntactic distinctions that are left unmarked in many other languages. The fact that many or most languages operate without marking the distinctions in question entails that to mark the feature at all is ornamental, rather than necessary, to human language. As such, in a given area, one language can be considered *overspecified* in comparison to another. (McWhorter 2005: 268; emphasis in the original)

Given that languages without affirmative markers are well attested, it might be said that languages with affirmative markers are overspecified in the area of polarity. Crucially, this is more than a rechristening. McWhorter (2005) understands the concept as a gradient such that languages that have been spoken uninterruptedly for millennia, but have never had a notable number of L2 speakers, will show a relatively stronger bent for overspecification. It is those languages, then, that are particularly likely to have overt structures for functions that are left implicit in most other languages (including creoles); cf. also Berdicevskis & Semenuks (2020: 291, 293; 2022: 18).

The basic idea underlying the notion of overspecification is thus that some languages come closer to exhausting their morphosyntactic options than others. This premise has also featured in other theoretical accounts, and it should be pointed out that the terminology is less critical to the present work than the content. As such, one might also refer to overspecification as ‘verbosity,’ which Dahl (2004: 54) defines as the “cross-linguistic dispensability” of a category or meaning and which he himself considers essentially equivalent to overspecification (Dahl 2004: 80, fn. 2). More generally, one might subsume the concept under the phenomenon of ‘informativeness,’ where the idea is that highly frequent values (such as ‘affirmative’ within the category of polarity) are less communicatively informative than infrequent ones. Regardless of the terms, the strong prediction based on virtually all previous

⁵ Note that McWhorter’s approach to creole languages has engendered considerable controversy (e.g., DeGraff 2003; cf. also Bisnath et al. 2026). However, this is of no concern here since there are no creoles in my sample and no claims about creoles will be made. Rather, McWhorter’s concept of ‘overspecification’ simply happens to also be useful if taken in isolation, and I consider the adoption of an existing term preferable to coining a synonym. Cf. also below.

work in morphosyntactic typology is that languages will prefer to encode the less frequent and subsequently more informative functions overtly. It follows that the expression of the more frequent values due to overspecification/verbosity, wherever it might occur, is a phenomenon of typological interest.

If overspecification is a tendency that emerges from long periods of exclusive L1 acquisition, it should result in a rather large number of grammatical paradigms overall. Couched within a grammaticalization approach, the prediction would be that languages that create relatively many categories over time also tend to express the exponents of those categories by means of bound morphology. To the extent that that is the case, the profile of overspecifying languages might be '(poly)synthetic' in traditional terms. While there is nothing in McWhorter's (2005) account or in the present approach that limits overspecification to the morphological (as opposed to the syntactic) level, it will be seen in Section 3 that the morphological domain is indeed the relevant one for affirmative markers.

The evaluation of the hypothesis that affirmative markers cluster in overspecifying languages will be at the heart of Section 4, where I will also describe how the concept of overspecification is operationalized in this work. As is the case with all typological predictions, however, the present one will need to be understood in probabilistic terms. That is, the hypothesis is not that it would be possible to classify any potential language with affirmative markers as overspecifying. Similarly, no overspecifying language will necessarily have affirmative markers. More generally, the notion of overspecification is only one way to approach affirmative markers. It is chosen here as an explicitly tentative explanatory angle, mainly because no other theoretical account suggests itself. As such, the present work will hopefully inspire further research on the factors that lead to the emergence of rare types of markers.

To sum up, affirmative markers would be of interest to typology because their overarching category is binary, and zero-marking the affirmative function is unlikely to be misinterpreted given that the only value it contrasts with is always overtly marked. Hence, the category of polarity is sufficiently specified with only negators, and the baseline expectation is that languages will not also have ('uninformative') affirmative markers (cf. Diessel 2019: 228). Conversely, languages that do have affirmative markers would be overspecified in the above sense. In conjunction with their presumably high token frequencies, affirmative markers would thus be compelling evidence for the idea that languages may prefer explicitness to economy, at least locally.

Section 2 is dedicated to the methodological decisions that shaped the database, the latter of which will be analyzed in Section 3. Meanwhile, Section 4 will discuss the results against the background of overspecification and general findings in morphological and diachronic typology. Section 5 will conclude and suggest avenues for further research.

2. Method and definitions

Cysouw & Wohlgemuth (2010: 5) cite the idea that linguistic phenomena that occur in 1-5% of the world's languages should be considered 'rara.' If 'affirmative' is indeed marked infrequently, the relevant morphemes might thus qualify for *rarum* status on this (admittedly arbitrary) criterion. This, however, would also mean that affirmative markers cause the same empirical issues as other *rara*: they cannot be searched for systematically and require large samples as well as a laborious data collection process (Cysouw & Wohlgemuth 2010: 5). In view of this situation, I searched for affirmative markers in reference grammars of 604 languages from 179 families (where each isolate counts as a family), as defined by Glottolog (Hammarström et al. 2026).⁶ Strictly speaking, this collection constitutes a convenience sample because it is made up of the languages for which I had access to grammatical descriptions. Yet, I am not aware of any serious biases in that sample, and it is hoped that its sheer size renders it more reliable than the average convenience sample. The 179 families (including isolates) represented in my sample correspond to 41.6% of the 429 families (including 183 isolates) listed in Glottolog as of this writing.⁷ Table 1 lists the number of families and languages surveyed for this project, sorted by the six geographical macro-areas that underlie the classification in Glottolog. The actual languages and families are provided in the Appendix.

In each grammar, I performed keyword searches (where possible), scoured the table of contents (sometimes complemented by the subject index) for all potentially relevant passages, and scanned the sections on verbal morphology as well as any information about particles and clitics.

⁶ In the interest of consistency and comparability, I also rely on Glottolog for all names of languages and families mentioned here, including where they diverge from the glossonym employed by the relevant sources and/or where they might be rejected by the speech communities at issue.

⁷ <https://glottolog.org/glottolog/glottologinformation> (last access: June 22, 2026).

Macro-area	Number of families represented	Number of languages surveyed
Africa	28	101
Australia	16	51
Eurasia	25	134
North America	33	113
Papunesia	40	120
South America	37	85
Total	179	604

Table 1: Number of language families in sample, sorted by macro-area.

In addition to the reference grammars, I performed less thorough searches in about 75 sketch grammars and about 90 books (edited volumes and monographs) on specific language families or areas, some of which contained sketch grammars themselves. Few of these additional resources described language families not covered by the grammars, and none yielded novel data. Therefore, these languages are not part of the above-mentioned sample, and, with one major exception to be discussed in Section 3.1, the data and information in the next section thus come from reference grammars.

While the concept of negation covers a range of functions, the markers aimed at here were only those that are the semantic opposite of “standard negators” as defined by Miestamo (2005). These markers, like German *nicht* or Spanish *no*, occur in declarative clauses, have predicate scope, and are not inherently emphatic. The restriction to standard declarative negation rules out noun phrase negators (*no idea*), negative pronouns (*nobody*), and negative adverbs (*never*); cf. Dryer (2005: 454). Meanwhile, the condition of predicate scope excludes negators with lexical scope (as in *unkind*). On the basis of that definition, it follows that the affirmative markers of interest will also have to occur in, and have scope over, declarative predicates, while not being emphatic. Their main function is thus to present the predicate of the clause as factual. While the overt expression of such a notion might appear unusual, the affirmative value itself falls out from the semantic structure of a polarity paradigm in the same way that ‘third person’ falls out from a person paradigm, etc.

In examples (1a) and (1b), from Anal (anm; Sino-Tibetan, Kuki-Chin), the kind of affirmative marker relevant to this study is illustrated by contrasting it with the corresponding negator. The elements at issue here and throughout will be highlighted

in boldface. It can be seen that the clauses are minimal pairs, contrasting only in terms of polarity, and that the two exponents occur in the same post-verbal slot. Further details of this affirmative marker will be discussed in Section 3.1.⁸

(1) Anal (Sino-Tibetan, Kuki-Chin; Ozerov 2018: 715)

- a. *tùŋ-kà-hín-nín*
dry-AFF-PL-1
'We (EXCL) are dry.'
- b. *tùŋ-má-hín-nín*
dry-NEG-PL-1
'We (EXCL) are not dry.'

Given that polarity is binary, the most straightforward situation would be for affirmative markers and negators to form a polarity paradigm whose members are mutually exclusive (cf. Malchukov 2011: 263). In such a scenario, affirmative markers would occur in exactly those clauses in which standard negators do not appear, and each clause would feature the affirmative marker or the negator. Yet, as will be seen in the next section, the occurrence of affirmative markers is usually subject to further constraints, and so a perfectly complementary distribution of the two exponents is rarely found. As such, the most important distributional criterion for the present study is that affirmative markers and negators cannot co-occur at the clause level. This is because no clause can predicate something and simultaneously negate that same predicate. If a negator and an affirmative marker are claimed to occur in the same clause, (at least) one of the two markers must actually perform another function (cf. Vázquez Álvarez 2011: 296–297 on Chol⁹ for one such case). Any instances of putative affirmative markers occurring alongside negators were disregarded and not counted toward the database. Finally, the focus on declaratives simply entails that affirmative markers *can* occur in that clause type; it does not mean that affirmative markers are *restricted* to declaratives. In fact, affirmative markers that occur in interrogative

⁸ I will homogenize glossing abbreviations across sources and examples. In particular, I will gloss all elements analyzed as affirmative markers as 'AFF.' Beyond that, I will follow the Leipzig Glossing Rules (Comrie et al. 2015) to the extent possible.

⁹ Chol; ctu; Mayan, Western Mayan.

clauses crucially aid in distinguishing affirmative polarity from declarative mood (cf. below).

Despite the rigid criterion of mutual exclusivity, it needs to be mentioned that explicit information or evidence concerning the compatibility of negators and alleged affirmative markers is not provided in every case. Thus, some of the analyses below might be subject to debate. Furthermore, a strict paradigmatic organization of the affirmative marker and the corresponding negator(s) would mean that the former cannot be optional because standard negators are not optional either (Dahl 2004: 137). As will be seen, however, this criterion is violated by some of the otherwise credible affirmative markers.

As suggested above, there are certain functions that an affirmative marker might be confused with. First and foremost, the affirmative function needs to be distinguished from that of the declarative. The two will often co-occur given that affirmative declaratives are the most frequent kind of clause (Velupillai 2012: 346), but a declarative clause can be negated, unlike an affirmative clause. In addition, a declarative marker is likely to be in competition with other speech act or mood markers such as interrogatives, whereas the only semantic contrast for an affirmative marker is a (set of) negator(s).

Having said that, the data below will show that there seem to be cumulative affirmative/declarative markers. Similarly, affirmative is not equivalent to indicative, the latter of which should contrast with subjunctive markers (etc.) and also be compatible with negators. Analogous arguments hold for other typically zero-marked values in the verbal domain, such as realis mood. All of these categories are semantically compatible with negation and most likely stand in paradigmatic opposition to exponents other than negators. So, while the semantic contribution of declarative, indicative, and realis markers is difficult to pinpoint in isolation, they crucially differ from affirmative markers on distributional criteria, which will thus be invoked in the analyses below. Only once these functions are separated out does the idea that affirmative markers are rare truly suggest itself. In fact, Miestamo (2007: 305) questions the existence of affirmative markers because every item described as such in his sample is either emphatic or an expression of indicative mood. This shows that such adjacent functions are highly relevant to the present work, and the case studies below will discuss them to the extent that the sources permit.

With respect to the formal dimension, affirmative markers had to be grammatical items, incapable of acting as utterances on their own. Hence, affirmative markers as

understood here are not particles corresponding to ‘yes,’ nor interjections, question tags, or ‘positive polarity verbs’ (cf. Mihas & Pérez 2015: 562 on Ashéninka Perené).¹⁰ Another way to think of the phenomenon is that affirmative markers will resist any attempt at translation into English, even by means of circumlocution, since English does not have standard (i.e., non-emphatic) affirmative forms (cf. [1a]).

It should be clear from the discussion so far that the present work will only deal with dedicated affirmative markers, i.e., elements that explicitly express the function of ‘affirmative.’ This is to distinguish them from well-known morphosyntactic effects that polarity distinctions might have elsewhere in the clause, especially in the tense-aspect-mood (TAM) system. That is, the fact that there is often a more limited range of TAM functions in negated than in affirmative clauses¹¹ does not mean that the additional TAM markers available in affirmative clauses are affirmative markers themselves. Note that by focusing on elements that express ‘affirmative,’ I will not always be able to determine from the description whether that is the only function of the relevant unit. In light of this, I will cast a rather wide net and also include certain items for which it is stated that ‘affirmative’ is one of their functions. Details will be discussed in the relevant empirical sections. On a general level, such an inclusive approach seems appropriate in an exploratory study given that it will be easier to discard false positives than to retrieve false negatives.

One domain in which both polarity functions may be marked is that of copulas, which often differ (suppletively) in affirmative and negative contexts. Affirmative copulas will not be considered affirmative markers for the purposes of this project because standard negation, the benchmark for the definition of affirmative markers, has a verbal predicate (Miestamo 2005: 1) and thus lacks a copula (i.e., a semantically bleached item linking to non-verbal predicates). Beyond that, affirmative markers of all overt kinds were considered for this project; that is, there was no a priori restriction to morphological and syntactic units. So, tonal affirmative markers, for example, would have been included. Yet, one of the main conclusions from the database is that all affirmative markers are segmentally instantiated affixes (or clitics).

¹⁰ Ashéninka Perené; prq; Arawakan, Southern Maipuran.

¹¹ This effect is itself a manifestation of markedness as traditionally conceived. Greenberg (2005 [1966]: 29) calls it “defectivation.”

3. Affirmative markers

In this section, I will discuss items that can be analyzed as affirmative markers given the description in the source. However, the amount of information and exemplification provided for the (putative) affirmative markers is never extensive. Therefore, not all of the criteria discussed in Section 2 can reliably be confirmed for every instance, even crucial questions will occasionally remain open, and the individual sub-sections are rather brief. Note in particular that the absence of affirmative markers from negated clauses cannot be demonstrated by any finite number of examples. Hence, statements concerning such an inability in the relevant sources will be taken at face value. Overall, the case studies presented below form a gradient from more to less straightforward instances, and the first three cases are the most convincing ones. Section 3.10 will provide an interim summary.

3.1. *Anal (anm; Sino-Tibetan, Kuki-Chin)*

Information on the Anal affirmative marker comes from Ozerov (2018, 2019) and Pavel Ozerov (p. c.). Note that Anal is the only language discussed here for which the facts and data were not drawn from a reference grammar. Since Ozerov (2018) is dedicated to tone and Ozerov (2019) discusses person indexation, information on the affirmative marker is generally sparse. However, Pavel Ozerov is the primary fieldworker on the language, and he confirmed that the affirmative marker, which he had independently analyzed as such previously, meets both the semantic and formal criteria stipulated in Section 2.

The affirmative marker in question is the suffix *-ka*, whose interaction with tonal marking will be ignored here. Evidence for the idea that this suffix marks the ‘affirmative’ function is copious: it only occurs in affirmative contexts (cf. Ozerov 2019: 37), appears in the same slot as the negator (Ozerov 2019: 40), and is thus mutually exclusive with it. Those aspects were discussed for examples (1a) and (1b) above. In addition, the affirmative marker can occur in interrogative verb forms (cf. Ozerov 2019: 40) and thus does not show the behavior of a declarative marker. Note further that an inability to occur in negated constructions not only argues against a declarative analysis but poses a similar problem for the idea that the element might be an indicative marker. Both declarative and indicative markers should be compatible with negators. On the other hand, it should be mentioned that the

affirmative marker does not occur in all affirmative contexts. Specifically, it is absent from transitive non-future clauses, where the affirmative function has been taken over by a nominalization construction (Ozerov 2019: 37, 39–40, 49). Example (2a) shows the affirmative suffix in a transitive future construction, while (2b) provides crucial evidence for its occurrence in an interrogative clause.

(2) Anal (Sino-Tibetan, Kuki-Chin; Ozerov 2019: 40)

- a. *abũ-tçú-to* *tçá-ná-kà*
 tree(type)-fruit-ABS eat-FUT-AFF
 ‘They will eat fruits.’
- b. *dá:²* *à-:-pé:-ká-nín*
 what 2-NON.AG-give.FUT.SAP-AFF-1
 ‘What shall I give you?’

Idiosyncratic restrictions such as the inability to occur in transitive non-future contexts will be seen for each of the markers discussed in this section. Rather than trying to account for them in the individual case studies, I will make an attempt at a general explanation in Section 4.2. There, I will also address the tentative diachronic link that Ozerov (2019: 49) suggests for the affirmative marker in Anal and the copula in closely related Monsang Naga (nmh; Sino-Tibetan, Kuki-Chin). For the time being, the evidence for an affirmative analysis of the Anal suffix is straightforward on both formal and functional grounds. As such, the suffix is the most compelling piece of counterevidence to the aforementioned ideas that affirmative markers either do not exist at all or are declarative/indicative markers instead.

3.2. *Isimjeega-Rootigaanga* (–; Nilotic, Southern Nilotic)

In *Isimjeega-Rootigaanga*, the affirmative and the negative prefix occur in the same slot of the verbal structure (Griscom 2019: 122) and therefore tend to be mutually exclusive – with the exception of some complicating factors concerning auxiliary constructions and their lexicalization (Griscom 2019: 186). Both the function and the distribution of the affirmative marker seem straightforward given that the “affirmation of a proposition is typically coded by the affirmative prefix” (Griscom 2019: 187). Furthermore, while the prefix does not appear in the imperative (Griscom

2019: 188), it is clearly not a declarative marker because it does occur in interrogative verb forms (Griscom 2019: 75). Generally speaking, the absence of any marker from imperative verb forms is unremarkable since imperatives are commonly bare forms (cf. Becker 2024: 397). Therefore, the inability of affirmative markers to occur in imperatives is not a counterargument to an affirmative analysis. As far as distributional restrictions are concerned, the affirmative marker is barred from conditional contexts (Griscom 2019: 188).

Example (3a) shows one of the two allomorphs of the affirmative marker, which are conditioned by a complex set of principles (Griscom 2019: 187). Meanwhile, example (3b) illustrates the occurrence of the affirmative prefix in an interrogative verb form. This form is actually marked for ‘affirmative’ in two different slots, once by each prefix allomorph (cf. Griscom 2019: 188 for the details).

(3) Isimjeega-Rootigaanga (Nilotic, Southern Nilotic)

- a. *q-à-bár ɲàn bòn-ér-d*
 AFF-3-hit earth people-PS.SG-SS.SG
 ‘People farmed.’ (Griscom 2019: 188)
- b. *g-àd-q-à-más*
 AFF-PSI-AFF-3-agree
 ‘They still agree?’ (Griscom 2019: 75)

There is little additional information about the affirmative prefix in the source. Overall, though, this is one of the most convincing cases discussed here. The relevant marker is not only semantically compatible with an affirmative analysis but is also in (more or less) complementary distribution with the negator. Taken together, this is strong support for the idea that these two items express opposing values of a single polarity paradigm.

3.3. Kabardian (*kbd*; Abkhaz-Adyghe, Circassian)

Colarusso (1992: 125) states that Kabardian “appears to be unique in the world in having a distinct mood mark for simple positive declaratives (in all but the present active tense).” Apart from the fact that Applebaum (2013: 60) provides a different analysis of the distribution of this item, this claim requires some clarification. First,

it can be seen throughout this section that Kabardian is not in fact unique in this regard. Second, the affirmative marker is apparently described as an exponent of mood because it is limited to declarative clauses. Therefore, the question arises as to whether it is more properly analyzed as a declarative or an affirmative marker. The strongest argument for classifying it as an affirmative marker comes from Applebaum (2013: 62), who states that the ‘declarative’ marker (which is also the term she uses) cannot co-occur with negators. It follows that the affix in question only occurs in affirmative contexts, which are therefore the delimiting factor. As argued above, if this item were a declarative marker, its inability to occur in negated clauses would not be accounted for. It should also be noted here that the negator and the affirmative marker occur in the same suffix slot (Applebaum 2013: 60), and the resulting incompatibility of the two markers further argues for an affirmative and against a declarative function. Additional evidence for the affirmative analysis comes from the Russian literature, which has traditionally described this item as an affirmative marker (e.g., Kumakhov 1989, 2006).¹²

However, the suffix does not occur in interrogative clauses (Colarusso 1992: 125, 126), which might support the declarative analysis. The bottom line is that every clause in which the suffix appears is affirmative, while the suffix does not occur in all declarative contexts given its absence from negated clauses. It thus seems defensible and more descriptively economical to consider it an affirmative marker, though with a potentially significant declarative component. Such an interaction is generally plausible given the frequent overlap of those two functions, and further evidence for this idea will be provided in following sub-sections.

Examples (4a) and (4b) illustrate the relevant affirmative item in the Russian and in the Turkish variety, respectively. Note that Colarusso (1992: 125) remains non-committal with regard to the exact shape of the suffix.

(4) Kabardian (Abkhaz-Adyghe, Circassian)

- a. *sə-λaaʒa-ś(a)*
1SG-work-AFF
‘I work.’ (Colarusso 1992: 125)

¹² I am deeply indebted to Ludger Paschen for finding and summarizing the Russian sources for me.

- b. *zə ʈɒpq ʃə-ʔv-s*
 one tribe OBL-exist-AFF
 ‘There was a tribe.’ (Applebaum 2013: 197)

3.4. *Bambassi (myf; Blue Nile Mao)*

M. Ahland (2012: 387) states that his analysis of the affirmative prefix in Bambassi is ‘tentative’ because the item has a ‘perplexing’ distribution. Among other things, it is optional and seems to interact with mood, clause type, and person (M. Ahland 2012: 390–398). Crucially for the present analysis, the marker is prohibited in all negative contexts, which is a necessary condition for affirmative markers but not for declarative and/or indicative markers. It can also occur with certain interrogatives, which suggests that it is not a declarative marker (M. Ahland 2012: 387, 388). Interestingly, the prefix had previously been analyzed as a declarative marker, but this runs counter to its incompatibility with negative contexts (M. Ahland 2012: 388). This shift in analysis also gains support from the fact that there is at least one actual declarative affix in the language, which can be seen in (5a). The parentheses in that example are taken over from the source, where they presumably signal the optional status of the affirmative prefix. Meanwhile, example (5b) illustrates the use of the affirmative marker in a question.

(5) Bambassi (Blue Nile Mao)

- a. *(ha-)tí-pè:f-á*
 (AFF-)1SG-slap-DECL
 ‘I slapped (it).’ (M. Ahland 2012: 390)
- b. *íʃ mùnts’-ìʃ hì-àn ha-hój-à:*
 DEF woman-SBJ 2SG-COM AFF-go-INTR
 ‘Did the woman go with you?’ (M. Ahland 2012: 474)

Based on the description by M. Ahland (2012), there is no major argument against the affirmative analysis of the Bambassi prefix, but of course the marker is insufficiently understood. That said, its optional status and restriction to certain

contexts are potentially indicators of its recent grammaticalization.¹³ If so, this might suggest that affirmative markers grammaticalize after negators, which would be in accordance with the notion of overspecification outlined in Section 1. This issue will be discussed further in Section 4.2.

3.5. *Laz (lzz; Kartvelian, Zan)*

There are several affirmative ‘preverbs’ in Laz. These will be treated as prefixes here, following Lacroix (2019: 840), who shows that they have fixed slots in the verbal structure. These markers never occur in negated clauses, relative clauses, with non-initial verbs in cases of coordination, or with certain middle verbs (Lacroix 2009: 433, 437; 2019: 843). In addition, they are sensitive to information structure and involved in TAM marking (Lacroix 2019: 844–845). The inability of the markers to occur in negated contexts lends credence to the affirmative analysis, whereas the other restrictions at least do not argue against it. Note that Lacroix’s (2009, 2019) own analysis of the matter is equivocal; this will also be reflected in the present case study.

While Lacroix (2009) describes in detail which of the affirmative markers are chosen with which kinds of verbs, that choice seems to depend on the semantics of the surrounding verb rather than on that of the affirmative items themselves. I will thus limit myself to an illustration of the affirmative marker *do-*, for which the source provides the most clear-cut evidence of affirmative usage. None of the other ‘preverbs’ are claimed to be affirmative elements here. The prefix *do-* occurs with the majority of verbs that do not also bear a spatial preverb (Lacroix 2009: 435). The only example that explicitly shows the (possible) affirmative function is reproduced in (6). Note that both the glossing and the translation in the original source are in French and were translated into English by the present author.

(6) Laz (Kartvelian, Zan; Lacroix 2009: 436)

badi-k ncumu-ši oxoi do-k’od-u
old.man-ERG salt-GEN house AFF-build-AOR.3SG
‘The old man built a house of salt.’

¹³ This assessment especially suggests itself on the basis of the short texts in M. Ahland (2012: 713–725), where the affirmative marker fails to occur in many of the contexts where it would seem to be semantically eligible.

Overall, then, the claim that the above prefix is an affirmative marker is not entirely secure, even though Lacroix (2009: 438) emphasizes that the incompatibility of the affirmative items with negative contexts is one of the few traits that the prefixes share. He also states that they are very frequent (Lacroix 2009: 433), which should be the case given the high token frequency of affirmative clauses. On the other hand, he highlights that there is a single example in his corpus where an allegedly affirmative preverb occurs in a negated clause, specifically in a prohibitive construction (Lacroix 2009: 438). While this would be a sufficient reason to discard the element from the current database, the prefix in the prohibitive construction is not the one seen in (6). Hence, there is no strong evidence to suggest that *do-* violates the definition of an affirmative marker used here. Clearly, though, a better understanding of the many restrictions to which it is subject might eventually challenge that analysis.

Lacroix (2019: 844) calls the affirmative preverbs “one of the most intricate issues in the grammar of Laz.” Analogous statements are also found for some of the other languages (cf. Bambassi above), which suggests that the contribution of (putative) affirmative markers is cognitively challenging. Perhaps, however, this impression mainly emerges because languages with affirmative markers tend to be described by grammarians whose native languages do not have them.

3.6. Northern Gumuz (*guk*; *Gumuz*, *Nuclear Gumuz*)

The initial affix in the Northern Gumuz verb structure is called an ‘affirmative’ marker because it does not occur with interrogatives or negatives (C. Ahland 2012: 206, 209). While the former restriction would be compatible with a declarative function, the latter suggests that the element is more likely to be a cumulative affirmative/declarative marker. In addition, the item is not found with all affirmative verbs in complex clauses (C. Ahland 2012: 206). Note that C. Ahland (2012) refers to the category that the affirmative prefix instantiates as ‘mood,’ but this is presumably because of the heterogeneous elements with which it competes for the initial verb slot, i.e., subordinate markers and the jussive (C. Ahland 2012: 207). Example (7) illustrates the affirmative prefix.

(7) Northern Gumuz (Gumuz, Nuclear Gumuz; C. Ahland 2012: 207)

d-á-gáχ

AFF-3SG.INTR-be.good

‘It was good.’

Since the function of the affirmative marker is not further elaborated upon in the source, the analysis provided there will be accepted as is. It is also of interest that Northern Gumuz has an affirmative affix but no negator affixes (C. Ahland 2012: 184). On the face of it, this would seem to contradict the development hinted at in Section 3.4, according to which affirmative markers grammaticalize after negators. On the other hand, it might of course be that negators in Northern Gumuz have been subjected to more erosion and/or renewal than the affirmative item, perhaps as part of the well-known negative cycle (van Gelderen 2008). To the extent that the current negator has only been established recently, this would help to explain its lower degree of fusion. This general point will be discussed in Section 4.2.

Interestingly, the affirmative prefix in Southern Gumuz (-; Gumuz, Nuclear Gumuz), which is a separate language according to Glottolog, can occur in questions. This element is segmentally different from the one in Northern Gumuz but occurs in the same verb slot (cf. C. Ahland 2012: 206). The prefix in an interrogative verb form is shown in (8).

(8) Southern Gumuz (Gumuz, Nuclear Gumuz; C. Ahland 2012: 208)

b-á^h-tsá-gá = áj

AFF-2SG.INTR-go-N.FUT = Q

‘Did you go?’

This distribution suggests that the Southern Gumuz item is less of a declarative and more of an affirmative marker than the one in Northern Gumuz shown in (7). However, the description by C. Ahland (2012: 207) suggests that the interrogative usage in (8) is exceptional, and hence this is difficult to evaluate. I will return to this matter as part of the diachronic discussion in Section 4.2.

3.7. Karo (Ethiopia) (*kxh*; South Omotic)

In Karo, there is a marker of affirmative/declarative function that Belete (2018: 177) calls a suffix. However, since it can occur on verbal and non-verbal predicates alike, it might best be analyzed as a clitic instead (Zwicky & Pullum 1983; Spencer & Luís 2012; Zingler 2022). More importantly, the claim that the element marks the declarative function is based on the fact that it never co-occurs with imperative, hortative, or interrogative morphs (Belete 2018: 179). It also does not occur in

dependent clauses (Belete 2018: 179). Meanwhile, there is another postposed element with the same restrictions that occurs in negated declaratives following the negator affix (Belete 2018: 179). In sum, then, the dichotomy between the affirmative/declarative marker and the negative/declarative marker seems straightforward. As such, the Karo marker is more clearly cumulative than the ones in Kabardian and Northern Gumuz sketched above. The former element is illustrated in (9a), and its negative counterpart is given in (9b).

(9) Karo (Ethiopia) (South Omotic)

- a. *wuti zoot-ite-na wo = do*
 1PL swim-PROG-AFF.DECL 1PL = exist
 ‘We are/were swimming.’ (Belete 2018: 177)
- b. *na ni-ma-mo*
 3SG come-NEG-NEG.DECL
 ‘She will not come.’ (Belete 2018: 178)

The marker in (9a) is arguably the best evidence for the close relationship between the affirmative and the declarative function. That is, while this interaction was already brought up for some of the other examples, none of the languages discussed so far expresses these overlapping categories as systematically as Karo. In addition, it might be cumulative elements like these that limit the distribution of pure affirmative markers (e.g., if the cumulative markers become pure declarative rather than pure affirmative markers). This will be taken up in Section 4.2. At the same time, the two markers show the crucial difference between affirmative and declarative markers. If *-na* in (9a) were merely a declarative, it would also be expected in (9b).

3.8. Male (Ethiopia) (*mdy*; *Ta-Ne-Omotic*, *North-West Ometo*)

The situation in Male closely resembles that described for Karo above. There is a cumulative affirmative/declarative marker that is in complementary distribution with the negative/declarative marker, neither of which co-occurs with interrogative, optative, or imperative morphs (Amha 2001: 148). Here too, then, there is enough evidence to suggest that the marker expresses both ‘affirmative’ and ‘declarative.’ It is furthermore of interest that the affirmative/declarative marker is limited to

independent clauses and to what Amha (2001: 148) calls ‘neutral’ statements, which she distinguishes from emphatic contexts. This sufficiently proves that the marker at issue meets another functional requirement spelled out in Section 2, namely, the inability to be used emphatically. A major difference compared to Karo is that the marker always occurs on the verb, including when the verb does not appear clause-finally (Amha 2001: 149–150). Unlike the Karo element, then, this marker appears to be a bona fide suffix. The affirmative/declarative marker is shown in (10a), and its negative equivalent is exemplified in (10b).

(10) Male (Ethiopia) (Ta-Ne-Omoti, North-West Omoti; Amha 2001: 148)

- a. *ʔ-atsí* *ziginó* *mukk-é-ne*
 person-M.NOM yesterday come-PFV-AFF.DECL
 ‘The man came yesterday.’
- b. *ʔ-atsí* *máári* *mazz-uwá-se*
 person-M.NOM house.ABS build-IPFV.NEG-NEG.DECL
 ‘The man is not building a house.’

The obvious formal, functional, and distributional similarities between the affirmative markers in Karo and Male raise the question of whether this overlap has come about by chance. While language contact is one possible alternative, it should be pointed out that South Omoti and Ta-Ne-Omoti are both branches within the Afro-Asiatic family in the *World Atlas of Language Structures* (WALS; Dryer & Haspelmath 2013). This differs from the classification in Glottolog, where the two units are separate families. In this case, then, the WALS classification might offer a more straightforward explanation of the parallels between the two markers, but the present argument does not hinge on this issue.

3.9. Southern Yukaghir (*yux*; *Yukaghir*, *Kolymic*)

The Southern Yukaghir verb is not usually marked for positive polarity (Maslova 2003: 144). In other words, the affirmative prefix found in the language is optional. Beyond that, the marker also has distributional preferences in that it is used most frequently in the irrealis and can only occur with ‘main’ forms, which seems to mean those that are not subject focus or object focus forms (cf. Maslova 2003: 139, 144). An additional complication is that the marker can be used to express emphatic

affirmation (Maslova 2003: 144), which might violate a major criterion specified in Section 2. However, the examples provided throughout the grammar do not suggest that the marker generally expresses emphasis, and it will therefore be interpreted as capable of encoding the pure ‘affirmative’ meaning. One apparently non-emphatic example, in realis mood, is given in (11a). Meanwhile, (11b) is relevant for two different reasons. On the one hand, it shows that the item can occur in questions and is thus not a declarative marker. On the other hand, it co-occurs with an irrealis marker, thus illustrating the semantic difference between affirmative polarity and realis mood.

(11) Southern Yukaghir (Yukaghir, Kolymic)

- a. *me-met-ahurpe-š-čē*
 AFF-REFL-suffer-CAUS-1SG.INTR
 ‘I make myself suffer.’ (Maslova 2003: 145)
- b. *tiŋ korobka-ge m-et-šøw-jemet*
 this box-LOC AFF-IRR-enter-2PL.INTR
 ‘Would you enter this box?’ (Maslova 2003: 172)

Apart from the exact functional contribution, there is another factor complicating the analysis of the above element. The standard negator in the language is also a verbal prefix, but it is not assigned to the same slot as the affirmative marker (cf. Maslova 2003: 139), unlike in Anal or Isimjeega-Rootigaanga. Hence, there is at least the possibility that the affirmative marker and the negator can co-occur, which would critically undermine the idea that the former is an affirmative marker. Then again, Malchukov (2011: 264) points out that exponents of complementary functions can co-occur to produce non-compositional meanings. Such a scenario would presumably not affect the present analysis.

3.10. Interim summary

This section discussed nine plausible affirmative markers from as many languages. While I know of no major arguments to question the classification of the items in Anal, Isimjeega-Rootigaanga, or Kabardian, the other items are of a more dubious status. This is either because they are (possibly) not ‘pure’ affirmative markers (Northern Gumuz, Karo, Male) or because the functions of the putative affirmative

markers are generally somewhat unclear (Bambassi, Laz, Southern Yukaghir). Overall, though, this database is sufficient to show that affirmative markers exist in at least some languages. This result thus disproves Dixon’s (2012: 90) strong claim that affirmative markers are categorically absent from languages. As will be argued below, the database lends credence to the idea that cross-linguistic rarities should neither be unduly “normalized” nor unnecessarily “exoticized” (cf. Anderson et al. 2025: 8–9).

Going back to the notion of ‘rara’ introduced in Section 2, the database suggests that affirmative markers may indeed qualify for *rarum* status. The sample underlying this study contains 604 languages, of which no more than nine contain affirmative markers. This corresponds to 1.5% and would thus be safely within *rarum* range. It should also be noted that, if three or more of the less secure cases were discarded, the ratio would fall to or below 1%, thus making affirmative markers a ‘*rarissimum*’ on the same arbitrary quantitative criterion (Cysouw & Wohlgemuth 2010: 5). Given these uncertainties, it seems most prudent to forgo a strict definition of *rara* and/or *rarissima*. Rather, the present study demonstrates beyond doubt that affirmative markers are a minority option, and polarity stands a good chance of being the most lopsided grammatical category in terms of marking asymmetries. Another fact that underscores the rarity of affirmative markers is that they do not even seem to be common at the family level. That is, none of the languages that are in my sample and related to the ones discussed in this section are claimed to have affirmative markers (but bear in mind the problematic issues surrounding ‘Omotic’ and Gumuz). This suggests that affirmative markers arise only when a very specific set of conditions is met, which causes their distribution to be rather sporadic. Section 4.2 will spell out some ideas along those lines.

Table 2 shows the affirmative markers discussed in Sections 3.1–3.9, summarizing their form, type of exponence, and position relative to the head (i.e., the verb in every language except Karo, where the marker can also follow other word classes).

Language	Form	Type of marker	Position of marker
Anal	<i>ka</i>	Affix	Postposed
Isimjeega- Rootigaanga	<i>q/g</i>	Affix	Preposed
Kabardian	<i>ś(a)/s</i>	Affix	Postposed
Bambassi	<i>ha</i>	Affix	Preposed
Laz	<i>do</i>	Affix	Preposed

Language	Form	Type of marker	Position of marker
Northern/Southern Gumuz	<i>d/b</i>	Affix	Preposed
Karo (Ethiopia)	<i>na</i>	Clitic	Postposed
Male (Ethiopia)	<i>ne</i>	Affix	Postposed
Southern Yukaghir	<i>me</i>	Affix	Preposed

Table 2: Overview of affirmative markers.

As can be seen, all markers save one are affixes, and even the outlier is described as an affix in the source. This is noteworthy because free affirmative markers were also searched for. Equally remarkably, five of the eight affixes are prefixes. The relevance of these patterns will be discussed and substantiated in Section 4.2.

Another pattern that emerges from the present section is the geographical concentration of the languages with affirmative markers. All nine languages come from the macro-areas of Eurasia or Africa, as defined by Glottolog. Even more remarkably, more than half of the database derives from a salient cluster of languages in eastern Africa, while another possibly relevant niche is found in the Caucasus. This distribution is shown in Figure 1, created with the GlottoshinyR tool by Börstell (2021).

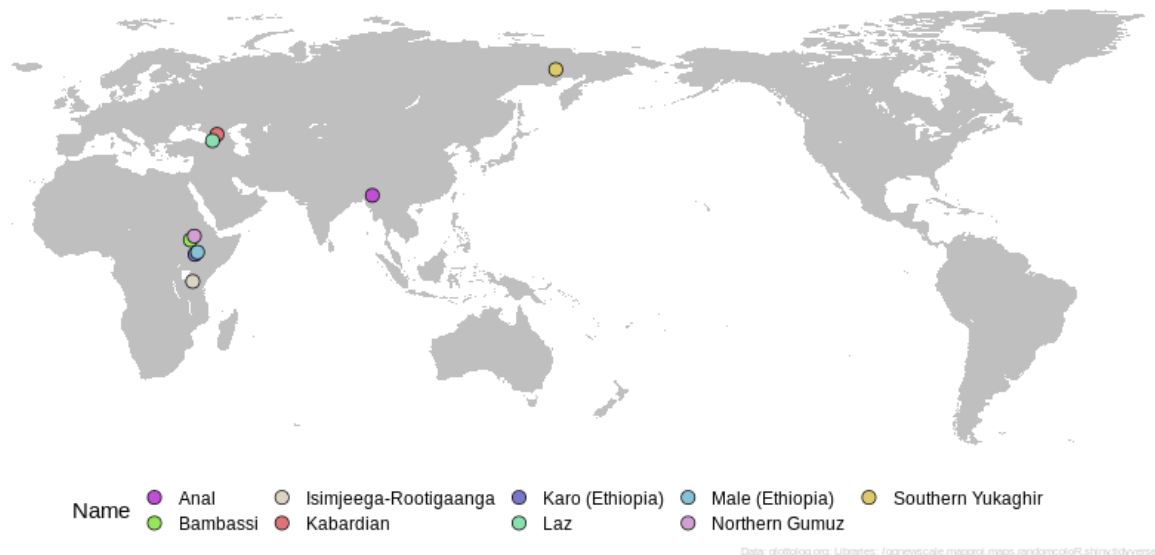


Figure 1: Geographical distribution of languages with affirmative markers.

Note that of the 179 families investigated here, only 53 are Eurasian or African (cf. Table 1). Hence, more than 70% of the families are from the other four macro-areas, and so the results cannot plausibly be attributed to a biased sample. While the geographical pattern suggests that affirmative markers might diffuse via contact, none of the markers likely to have been involved in borrowing for areal reasons are actually close in form, as can also be gleaned from Table 2 (this excludes Karo and Male, which are possibly related). The areal component will be taken up again in Section 4.1.

4. Discussion

This discussion aims to shed light on the empirical patterns described in Section 3. In so doing, it will focus on the question of why affirmative markers might occur in the languages in which they are found and why they show a slight prefixing preference. Note, however, that these proposals are necessarily speculative and do not claim to identify sufficient criteria for the existence of overspecification or affirmative markers. The alternative would be to skirt explanatory attempts entirely and to chalk up all *rara* to mere chance. In that sense, then, the tack chosen here is more ambitious and less fail-safe. The backdrop for the first part of the discussion will be McWhorter's (2005) concept of overspecification, which will be operationalized and applied in Section 4.1. Meanwhile, Section 4.2 will address major typological and diachronic issues raised by the data.

4.1. Affirmative markers as a product of overspecification

For the purposes of this discussion, overspecification will be broken down into two logically independent and falsifiable parameters: degree of synthesis and overt marking of other usually unmarked categories. Below, I will refer to the latter criterion as the 'systemic' parameter. Note that McWhorter (2005, 2007: 21–28) does not explicitly name these parameters as indicators of overspecification because he generally does not provide a rigid definition of this concept (cf. also Berdicevskis & Semenuks 2020: 284). Instead, the criteria that he employs are tailored to the specific languages that he investigates (often in pairwise fashion). However, the parameters suggested here are in tune with McWhorter's overall approach and mostly differ from

his in that they are more abstract so as to be applicable to a larger number of languages at once.

The first parameter concerns the degree of synthesis.¹⁴ The idea is that, if a language overspecifies, it will have a relatively large number of grammatical categories that it expresses synthetically. Synthesis is traditionally discussed with respect to the verb, and this (usually implicit) bias is presumably due to the fact that verbs tend to have richer morphology than nouns (e.g., Bybee et al. 1990: 4). However, potential effects of this bias are irrelevant to the present study since affirmative markers modify the meaning of the predicate and, as was seen above, are marked on the verb almost without exception. Hence, the empirical results suggest that, if anything, it is the verbal morphology that is part of the explanation.

The degree of synthesis in a language is often represented in reference grammars by means of a ‘template,’ which indicates the number of affix paradigms that can be manifested on a single finite verb form. Of course, such templates only indicate the logically possible maximum, and most verb tokens in languages with large templates do not approximate this maximum. Nevertheless, such templates are instructive when compared to languages like English, which has been argued to have a ‘monosuffix constraint,’ possibly the last stage before a fully analytic structure (Aronoff & Fuhrhop 2002: 485). Thus, a template for English verbs would be limited to a single inflectional slot, and templates in languages going beyond a single slot can thus be considered increasingly synthetic. It is this gradient understanding of synthesis that will underlie the analysis offered below.

I found explicit statements regarding the maximum verbal template for six of the nine languages that have affirmative markers, and I secured the template for Anal from Pavel Ozerov (p. c.). The two languages for which I do not have templates are the ‘Omotic’ ones, Karo and Male.¹⁵ Table 3 presents the number of prefix and suffix slots in the seven languages. In cases where the source gave several templates due to a difference conditioned by verb type, I chose the least marked one on the traditional definition. For Bambassi, this means the realis instead of the irrealis template, and for Northern Gumuz this means the non-future instead of the future template.

¹⁴ I will not deal with the potential difference between ‘synthesis’ and ‘polysynthesis,’ which is at best difficult to operationalize and possibly conceptually flawed (cf. Haspelmath 2018 for discussion).

¹⁵ I also did not find relevant statements/templates in any of the related languages for which I have grammars.

Language	Prefix slots	Suffix slots	Source
Anal	5	6	Pavel Ozerov (p. c.)
Isimjeega- Rootigaanga	3	5	Griscom (2019: 122, 126)
Kabardian	13	5	Applebaum (2013: 41)
Bambassi	2	7	M. Ahland (2012: 350)
Laz	4	7	Lacroix (2019: 840)
Northern Gumuz	7	6	C. Ahland (2012: 184)
Southern Yukaghir	3	4	Maslova (2003: 139)

Table 3: Number of affix slots in the verbal template of languages with affirmative markers.

The patterns in Table 3 are rather remarkable. The average number of prefix slots is 5.3, and the average number of suffix slots is 5.7. (If Karo and Male had no affixes at all, which is not the case, the average would be 4.1 for prefixes and 4.4 for suffixes.) As such, the average sum of prefix and suffix slots in the languages is roughly between 9 and 11, depending on the exact structure of Karo and Male. This puts languages with affirmative markers toward the highly synthetic end of the worldwide spectrum since Bickel & Nichols (2013) only find two languages with more than eleven templatic slots and only nine languages with more than nine templatic slots in their 145-language sample.¹⁶ Based on that evidence, the idea that affirmative markers are found in languages with a high degree of synthesis has been confirmed. More specifically, languages with affirmative markers not only show a strong bent for grammaticalization by creating exponents for a rare function in the first place, but also by expressing an above-average number of functions by means of bound morphology.

An additional aspect of interest is the relationship between prefixes and suffixes. On the one hand, there are essentially as many prefix slots as suffix slots on average, and on the other hand, there is not a single language without prefixes (though Karo seems to belong in that category; cf. Belete 2018: ix–xi). If Greenberg (1966: 92) is

¹⁶ Note that the classification adopted by Bickel & Nichols (2013) is not identical to the one used for the present project. In particular, on their counting procedure, English is assigned two inflectional slots because the verb can mark agreement and past tense (but not in the same word form). Hence, their approach seems to count paradigmatic options while mine counts syntagmatic options. Nevertheless, their study is sufficiently close in aim to the present argument, and it is the most suitable point of comparison that I am aware of.

correct in claiming that exclusively suffixing languages are fairly common, it can be concluded that languages with affirmative markers are (largely) drawn from the subset of languages that do have prefixes. The relevance of this finding will be addressed in Section 4.2.

The second parameter for the operationalization of overspecification is the number of other cross-linguistically zero-marked meanings that are overtly expressed in the languages with affirmative markers. The idea behind this approach is that, if overspecification is a general tendency within certain languages, most grammatical meanings in those languages might be encoded by overt material. I will not offer any ideas as to why such an effect might manifest itself in any given language except that analogy is one of the driving forces of language structure and change (e.g., Bybee 2010: ch. 4). That is, once a language overtly expresses one typologically unmarked function, this principle might be extended to other relevant meanings via a ‘transparadigmatic’ process (or, less formally, a ‘snowball effect’). Such an extension of overt marking might ultimately be triggered by a preference for ‘iconic’ structures, which supply every meaning with an overt form (cf. Wurzel 1987: 92; Croft 2003: 102; Cristofaro 2014: 289). Based on that reasoning and the improbability of affirmative marking, the systemic approach pursued here could also be phrased as a rough implicational hypothesis: a language that has affirmative markers will also have markers for other typically unmarked functions. That is the claim to be tested here.

In the interest of comparability, I will limit the investigation to the cross-linguistically common categories of case and number on nouns and tense and agreement (i.e., person marking, indexation) on verbs. For languages that do not have tense, I chose aspect. Based on the lists compiled by Croft (2003: 156–157) and Bybee (2015: 100), the typologically unmarked members of those categories are nominative and absolutive for case, singular for number, present for tense (or perfective for aspect), and first and/or third person (singular) for indexation. Hence, the prediction based on the systemic assumption is that languages with affirmative markers will also express (some of) those values overtly.

Table 4 lists the resulting patterns. Empty cells indicate either lack of evidence for non-zero marking or counterevidence to non-zero marking. All person markers are to be understood as singular. Where there are multiple paradigms of person markers, I chose the least ‘marked’ one on the traditional approach (e.g., indicative instead of subjunctive). The numbers in parentheses indicate the pages where the relevant information can be found in the corresponding sources used in Section 3. The page

numbers for Kabardian refer to Applebaum (2013). Information on Anal is from Ozerov (2019) or from Pavel Ozerov (p. c.).

Language	Number	Case	Person	Tense/aspect
Anal		Absolutive	First (27)	Present
Isimjeega- Rootigaanga			First (175)	
Kabardian		Absolutive (29)	Third (42)	Present (57)
Bambassi				Present (460)
Laz				Present (340–341)
Northern Gumuz		Nominative (147)	First, third (123– 126)	
Karo (Ethiopia)				
Male (Ethiopia)		Nominative, absolutive (56)		Perfective (114)
Southern Yukaghir			First, third (140)	

Table 4: Overt marking of typologically unmarked functions in languages with affirmative markers.

The results displayed in Table 4 are not as readily interpretable as those in Table 3¹⁷. In other words, a tendency to mark all members of a paradigm throughout the system is not as obviously a concomitant of affirmative marking as is verbal synthesis. First and foremost, there is no support from the category of number given that no language that marks affirmative also marks singular. This undermines the systemic idea in that number is relatively close to being a binary feature like polarity, unlike person and

¹⁷ Note the following complications with regard to Table 4. First, case in Kabardian is cumulated with definiteness: absolutive is unmarked when indefinite but marked when definite. Neither Croft (2003) nor Bybee (2015) list marking asymmetries for definiteness, and I did not find relevant statements in Greenberg (2005 [1966]) either. This is presumably because markedness values in the domain of definiteness are context-dependent (Givón 1991: 336). However, Becker (2024: 397) establishes that indefiniteness shows the strongest tendency to be zero-marked across the nominal, verbal, and adjectival categories she tested. If that means that ‘indefinite’ is the ‘unmarked’ value in a definiteness paradigm, the Kabardian pattern is unremarkable. Second, Kabardian also has zero allomorphs for third person and present tense. Finally, first person in Southern Yukaghir is unmarked for transitive indicative verbs but marked for intransitive indicative verbs; cf. example (11a) above.

case. On the other axis, Karo does not overtly mark any of the relevant values. As such, it provides counterevidence to the implicational hypothesis proposed above.

In addition, much of the positive evidence comes from person indexation, where the situation regarding coding asymmetries is unclear due to the general issue of whether first or third person is less marked. While Bybee (2015: 100) treats both persons as equally unmarked compared to second person, Croft (2003: 156) ranks third person as less marked than first. Hence, an evaluation of the category of person not only presents empirical problems but also theoretical ones. Finally, given that a language cannot have both a prototypical nominative and a prototypical absolutive due to functional overlap, it stands to reason that Male has at most one case marker supporting the overall idea investigated here. In sum, only the domains of case and tense/aspect provide any evidence for the systemic nature of overspecification.

There are some mitigating factors to explain these mixed results. Apart from the fact that Table 4 selected a subset of categories, the above analysis also introduced nominal morphology into the picture. Of course, it might be that languages only tend to overspecify the verbal domain, if any. That would at least line up with the fact that verbal morphology is usually more synthetic than its nominal equivalent. In fact, number and case are not marked at all in Isimjeega-Rootigaanga (Griscom 2019: 78), and Bambassi lacks case (M. Ahland 2012: 194). So, the comparison in Table 4 is not straightforward. Overall, 14 of the 36 cells in Table 4 are filled and 10 of the 18 cells for the two verbal categories. In order to assess whether these ratios are at all typologically meaningful, one would need analogous studies for different sets of languages and categories, but I know of no research addressing precisely this issue.¹⁸

In conclusion, it can be argued that synthetic verbal morphology is a necessary (if by no means sufficient) condition for the presence of affirmative markers, while the overt expression of other cross-linguistically unmarked functions is a lesser (though still suggestive) predictor. Thus, to the extent that overspecification was appropriately operationalized here, the hypothesis that affirmative markers tend to be found in overspecifying languages was broadly supported.

¹⁸ However, the worldwide marking patterns collected in Grambank (Skirgård et al. 2023) are still instructive in this regard. The data there show that present tense markers occur in 32.2% of all sample languages (771/2,392; GB082), singular markers occur in 17.3% of all sample languages (425/2,442; GB042), and core case marking on non-pronominal noun phrases, including nominatives and absolutes, occurs in 32.7% of all sample languages (777/2,377; GB070). Hence, all these unmarked functions are overtly expressed much more commonly than ‘affirmative’ is at 1.5%, which suggests that the systemic link between the former set and the affirmative markers might be very weak.

The idea that overspecification can spread across grammatical sub-systems also gains support from the fact that most of the languages with affirmative markers are spoken in eastern Africa. It is precisely languages from that area that commonly have marked nominatives (König 2006; Handschuh 2014: 221–223; cf. also examples [10a] and [10b] above). Of course, one would need to investigate if those items instantiate the prototypical definition of nominatives. Furthermore, there is the issue that languages with marked nominatives often have unmarked accusatives, so the presence of marked nominatives does not automatically entail overspecification. Nevertheless, the present study leads to the conclusion that the overt expression of typically unmarked functions is more likely to be found in eastern Africa than in other areas. Possible causes of this pattern will hopefully be the focus of future research.

As mentioned in Section 1, McWhorter’s (2005) argument is based on the assumption that overspecification is most likely in languages that have had few non-native speakers. As such, the sociohistorical profile of the languages in the database might provide ancillary support for the claimed link between affirmative marking and overspecification. While admittedly crude, the best way to approximate this issue might be to establish the number of speakers for each language. The idea is that small speech communities are more likely to be socially homogenous and/or geographically isolated. Therefore, the relevant languages are more likely to have had few L2 speakers and should thus provide the prerequisites for overspecification. So, the prediction is that languages with affirmative markers are spoken by small communities.

The following figures have to be understood against the background of the global trend of language loss. That is, the numbers given below may have greatly reduced by now. Apart from that, however, they seem to bear out the overall idea. Bambassi, Karo, Isimjeega-Rootigaanga, and Southern Yukaghir were each spoken by 5,000 people or fewer at the time or before the relevant grammar appeared (M. Ahland 2012: 13; Belete 2018: 17; Griscom 2019: 1; Maslova 2003: 19). Anal has 20,000 speakers (Ozerov 2019: 26), while Male has 46,000 (Amha 2001: 3). C. Ahland (2012: 3) gives 200,000 as the number of Gumuz people, which means that the number of speakers of (Northern) Gumuz might be considerably lower. Lacroix (2019: 830) concedes that the number of Laz speakers is unknown but cites figures as high as 250,000, whereas Applebaum (2013: 4) mentions 647,000 speakers of Kabardian.¹⁹

¹⁹ Daniel & Lander (2011: 125) give roughly similar numbers for the two Caucasian languages: “several dozens of thousands” in the case of Laz and “about one million” for Kabardian.

Overall, then, the sociohistorical aspect seems to tally with the idea that affirmative markers are an outgrowth of overspecification, at least if community size is an adequate operationalization.

4.2. Typological and diachronic issues

It was mentioned in Section 3 that the affirmative markers show a prefixing preference. This is unexpected on the assumption that they behave like morphological negators, which are more often suffixal than prefixal (Dryer 2013; Berg 2020: 376).²⁰ In fact, a juxtaposition of the affirmative markers and the standard negators in the corresponding languages shows that there is no symmetry in this regard. (As expected, all the languages with overt affirmative markers also have overt negators.) The relevant patterns are shown in Table 5.

Language	Type of affirmative marker	Type of standard negator
Anal	Suffix	Suffix
Isimjeega- Rootigaanga	Prefix	Prefix
Kabardian	Suffix	Prefix, suffix
Bambassi	Prefix	Suffix (multiple)
Laz	Prefix	Word (preverbal)
Northern Gumuz	Prefix	Enclitic
Karo (Ethiopia)	Enclitic	Suffix
Male (Ethiopia)	Suffix	Suffix
Southern Yukaghir	Prefix	Prefix

Table 5: Type of affirmative marker and standard negator.

Only six of the nine languages consistently position their affirmative markers and negators on the same side, and only four of the nine languages use the exact same type of form for their affirmative marker and their standard negator (unless the Karo

²⁰ Despite their overall suffixing preference, morphological negators are relatively more likely to be prefixed compared to affixal exponents of other categories. This is borne out by the data in Berg (2020) and Asao (2015: 24, 25). In both databases, indexation and possession are essentially the only domains that show a stronger proclivity for prefixation than does negation.

negator is an enclitic for the same reason that its affirmative marker was analyzed as one). At the very least, this positional variation seems to mean that affirmative markers and negators do not reliably show a strict paradigmatic organization. On the whole, these patterns argue against the idea that affirmative markers are created on the model of previously existing negators. Rather, as briefly suggested in Section 3.6, the possibly distorting effect of the negative cycle is not to be discounted here. That is, negators come and go more often than affirmative markers, and their age presumably impacts their degree of fusion. At the same time, degree of fusion also correlates with position in that grammatical markers generally show a more pronounced postposing bias on the morphological than on the syntactic level (Bybee et al. 1990; Asao 2015; Berg 2020). The heterogeneous pattern of the negators above is thus the more or less expected effect of the various forces that shape them.

Before the question of why the affirmative markers are predominantly prefixes will be addressed, it should be highlighted that this type of exponence is doubly unexpected. On the one hand, it is remarkable that the affirmative markers tend to be morphologically bound at all. In Berg's (2020: 376) sample, the negator is a free word in 300 languages and an affix in only 173 languages. This further illustrates the formal differences between negators and affirmative markers. The second discrepancy is the violation of the general suffixing preference. For instance, the worldwide 71-language sample in Bybee et al. (1990: 4) produced 1,236 verbal suffixes and only 426 verbal prefixes. Other things being equal, there should thus be about three times as many affirmative suffixes as prefixes in the database, contrary to the actual results. A juxtaposition of Berg's (2020) negator data and the affirmative markers in the present study yields a statistically significant difference at the 0.01 level on a Fisher's exact test, while an analogous test of Bybee et al.'s (1990) affix data and the affirmative markers is marginally significant at the 0.1 level.²¹

Based on the above, the main issue remaining is why affirmative markers show a prefixing bias. While the database might appear too small to speak of any patterns and to permit general inferences, it is unclear how much larger it could grow given the rarity of the phenomenon. Therefore, an attempt at explanation does not seem premature. One possible avenue is that the affirmative markers take the whole

²¹ Note that the Karo clitic required slight adjustments for both tests. For the comparison with the negator data, I classified it as a syntactic item, even though clitics combine properties of morphological and syntactic elements. In the second test, it was accommodated by translating Bybee et al.'s (1990) dichotomy between prefixes and suffixes into a difference between prefixes and non-prefixes.

relevant template as their model. That is, the affirmative markers might have become prefixes because the corresponding template already contained a sufficient number of prefixes (and/or prefix slots) to accommodate them via some kind of analogy (cf. Weir 1986: 314–315; Grossman et al. 2018: 533 for that line of argument; but cf. also Enrique-Arias 2002: 14–15 for the opposite claim). This idea is based on the reasoning that languages are more likely to build on the kinds of structures that they already have than they are to create new ones (cf. the concept of ‘system-congruity’ in Wurzel 1987: 62–76). This would be compatible with the idea that affirmative markers are formed predominantly once the process of overspecification is already under way, which, in principle, is supported by the patterns in Table 4. Further evidence for this chronological argument is that some of the affirmative markers are optional, which argues against an advanced stage of grammaticalization (Lehmann 2015: 148–150).

It should be mentioned that prefixes also emerge because of highly particular, language-specific factors (Mithun 2003; Grossman & Polis 2018). Yet, it is unclear why five different languages should all at some point have met the specific structural requirements for the formally and functionally unlikely development of affirmative prefixes. Thus, a more general explanation such as the one above seems preferable. On the other hand, the fact remains that the mere availability of prefixes would not *explain* why the affirmative markers became prefixes as well, not least because some of them actually became suffixes regardless. It follows that a full explanation of the distribution of the affirmative markers will need to be multifactorial, incorporating both typological and language-specific facts. Note that the focus here was on prefixation because the existence of affirmative suffixes can always be attributed to the factors responsible for the general suffixing preference (Bybee et al. 1990; Himmelmann 2014; Asao 2015; Berg 2020).

Another relevant point concerns the fact that four of the five languages with affirmative prefixes have more suffix slots than prefix slots, which dovetails with the finding that prefixes are also paradigmatically less elaborated than suffixes (Berg 2022: 392–393). One consequence of this difference is that prefixes are more predictable, which might support them even though they occur in an otherwise disadvantaged position for affixes (cf. Asao 2015: 121). The notion of predictability should be of even greater interest in the case of affirmative markers because the minimal semantic oppositions of which they are a part should render them highly predictable, i.e., less informative (cf. Section 1). Yet, since they do not necessarily occur in the same structural positions as the relevant negators and may even share a

slot with semantically unrelated markers (e.g., Northern Gumuz), it is not obvious that they are more predictable and/or easier to process than members of other paradigms.

The notions of overspecification and systemic analogy implicate the diachronic dimension in that they rely on the idea that (elements like) affirmative markers emerge as part of broader historical processes. Unfortunately, most of the languages with affirmative markers are not sufficiently documented to permit any strong claims about the history of the affirmative markers, and the relevant sources do not attempt any explicit reconstructions. It is also telling that the comprehensive *World Lexicon of Grammaticalization* (Kuteva et al. 2019) does not contain a single entry on affirmative markers, either as sources or as targets. As such, the kind of overspecification account attempted here largely has to rely on synchronic data, with all the pitfalls that follow from that method. Once that is accepted, though, some ideas about how affirmative markers might grammaticalize suggest themselves.

The data collection process for this project unearthed a number of elements that might be marginal affirmative markers but for which the description was insufficiently explicit.²² The most relevant set of markers of this sort subsumes the ‘particles’ in Southeastern Tepehuan (stp; Uto-Aztecan, Tepiman; Willett 1991: 158), Halh Mongolian (khk; Mongolic-Khitani, Eastern Mongolic; Janhunen 2012: 258), and Koasati (cku; Muskogean, Alabaman-Koasati; Kimball 1985: 459–460). These elements might lend credence to Bybee et al.’s (1994: 19–21) ‘Parallel Reduction Hypothesis,’ according to which grammaticalizing markers undergo semantic and formal reduction in lockstep. This is because these items, unlike the affirmative markers in Section 3, are either free or lightly bound, and they are also ‘heavier’ than the affirmative markers in terms of their function. For instance, the Southeastern Tepehuan particle still seems to carry some modal force, while the Koasati marker might still be emphatic to some degree.

A more diverse set of ‘near-affirmative’ items helps to shed further light on what kinds of meanings the affirmative markers might have grammaticalized from (assuming that there is at least occasional continuity between the near-affirmative

²² For instance, Sumerian (sux; isolate; Middle East) might have had two affirmative prefixes, but the descriptive situation is too uncertain to admit those items into the database (cf. Jagersma 2010: 577–579). I thank Victoria Birkner and Alexander Steiner for pointing me toward a decision on this issue.

and the actual affirmative markers).²³ For instance, in Abui (abz; Timor-Alor-Pantar, Central Alor; Kratochvil 2007: 280–281), a polysemous marker that may also express the ‘affirmative’ function is transparently derived from the verb ‘say,’ while a so-called ‘end-of-event’ marker in Mina (Cameroon) (hna; Afro-Asiatic, Chadic), which is at least implicitly affirmative, might derive from a verb/copula (Frajzyngier et al. 2005: 259). Another origin of near-affirmative markers is the particle ‘yes,’ which was indicated twice (Etxepare 2003: 538 for Basque²⁴; Linn 2001: 320 for Yuchi²⁵). Of course, none of this qualifies as definitive evidence, but some of these paths would be plausible. In particular, ‘yes’ is functionally compatible, while copulas resemble affirmative markers in their degree of semantic bleaching. Since rather many languages have affirmative copulas,²⁶ they might ultimately be the most likely origin (also recall the suggestion along those lines for Anal in Ozerov 2019: 49).

Another conceivable source of affirmative markers, which emerges from the Karo and Male data in particular, is cumulative declarative/affirmative markers, perhaps via the reanalysis from declarative via declarative/affirmative to affirmative. At a broader level, every diachronic assumption about affirmative markers made here suggests that they develop over long periods of time and are late additions to the morphological inventory of a language. In particular, the account sketched here lines up with the approach in Harris (2008, 2010) as well as with Dahl’s (2004: Ch. 6) concept of ‘maturation.’ Both suggest that the synchronic rarity of any given phenomenon owes to the fact that it only comes about after a complex (and thus long) sequence of specific diachronic stages.

Finally, it should be possible to tie the different morphosyntactic restrictions of the affirmative markers to their historical emergence. While some of the distributional patterns described in Section 3 are almost polar opposites (e.g., Isimjeega-Rootigaanga excluding affirmative markers from conditional clauses but Southern Yukaghir preferring them in irrealis contexts), there is also one low-level pattern that came to light. Specifically, several of the restrictions suggest that affirmative markers are most likely to occur in simple independent clauses. On the one hand, this might again argue for their relative recency given that subordinate clauses have been

²³ An anonymous reviewer asks whether evidential markers might be sources of affirmative markers. I did not find any indication of such a link, either among the clear-cut or the more dubious cases.

²⁴ Basque; eus; isolate; Western Europe.

²⁵ Yuchi; yuc; isolate; Oklahoma.

²⁶ In Grambank, 47.9% of all languages (1,152/2,407) have a copula with predicate nominals (GB117).

claimed to be more conservative (Bybee 2002; but cf. also Jing et al. 2023). On the other hand, the context of simple independent clauses would also render copulas a more likely origin than, for instance, verbs of saying, the latter of which are involved in complex sentences almost by definition.

5. Conclusion

On the basis of a large and balanced sample, this work furnished empirical evidence for the common claim that affirmative markers, the mirror image of standard negators, are highly infrequent in the languages of the world. At the same time, the few cases that were found are sufficient to undermine the more radical idea that affirmative markers are categorically absent from grammatical systems. The theoretical approach used to inform these data was McWhorter's (2005) concept of overspecification, understood here as the combination of a high degree of verbal synthesis and a comparatively low number of zero markers. This idea manages to account rather well for the data, but it only provides necessary rather than sufficient criteria. The database further suggests that affirmative markers show strong areal biases (eastern Africa), are confined to smaller speech communities, and grammaticalize rather late as well as from a variety of sources. Further research on all these aspects is necessary to establish why affirmative markers are found where they are and to the extent that they are. In addition, the results obtained suggest that the historical, functional, and formal links between affirmative and declarative markers merit further study, as do the syntactic/semantic restrictions of the former.

As extensions of this work, other typologically unmarked meanings would be obvious targets. Many of the relevant functions were mentioned or discussed above: declarative, active, indicative, realis, present, and perfective in the verbal domain; nominative, absolutive, and singular in the nominal domain (cf. also Mayerthaler 1987: 48). Based on the data collection for this work and the available data in Grambank, the prediction is that overt coding will not be as infrequent for any of these functions as it is for affirmative markers (no more than 1.5% of all sample languages). Meanwhile, a serious competitor for overt affirmative markers in terms of worldwide rarity might be morphemes expressing the basic, 'positive' degree of adjectives (e.g., *fast* rather than *faster* or *fastest*). Like 'affirmative,' this semantic value is rarely even thought of as a grammatical function, let alone as part of a paradigm, and neither Mayerthaler (1981: 27) nor Bobaljik (2012: 32, fn. 5) find a single

instance in their respective samples. Finally, it would be of considerable interest if creoles turned out to have affirmative markers. Based on the argument presented here, creoles are particularly unlikely to check that box given that affirmative markers only seem to arise in languages that already have large amounts of (old) morphology. However, that prediction is influenced by the specific database obtained here. That is, creoles might have syntactically free affirmative markers instead, which would then require a reconsideration of the reasoning pursued in this work. Future research will hopefully address these issues.

Assuming a broader perspective, further work on rarely marked functions would also permit a more profound assessment of overspecification. In particular, the causes of overspecification would be relevant to sub-fields beyond morphosyntax in that they seem to support the assumption that some systems might be more speaker-friendly (by requiring relatively little coding) while others might be more listener-friendly (by requiring more extensive coding). In that sense, the most interesting consequence of empirical research on such phenomena might be a new operationalization of a time-honored typological idea (cf. Bisang 2020 for some discussion).

Acknowledgments

This work benefited from the support of many people. Pavel Ozerov handed me the best example of an affirmative marker that I ended up finding and also provided crucial additional information on Anal. Ludger Paschen's selfless literature review greatly strengthened the case for Kabardian presented here. Thomas Berg and the audience at the 21st International Congress of Linguists in Poznań kindly let me know which parts of the argument to highlight and which to background. Victoria Birkner and Alexander Steiner generously discussed the complex Sumerian situation with me. Anonymous reviewers provided feedback that allowed me to make significant improvements with regard to content, structure, and argumentation. However, none of these scholars necessarily agree with the interpretation of the data or the framework adopted here, and any errors and infelicities are solely my responsibility.

Abbreviations

1/2/3 = 1st/2nd/3rd person
ABS = absolutive

AFF = affirmative
AOR = aorist

CAUS = causative
COM = comitative

DECL = declarative	LOC = locative	PROG = progressive
DEF = definite	M = masculine	PS = primary suffix
ERG = ergative	NEG = negation	PSI = persistive aspect
EXCL = exclusive	N.FUT = non-future	Q = interrogative
FUT = future	NOM = nominative	REFL = reflexive
GEN = genitive	NON-AG = non-agentive	SAP = speech act participant
INTR = intransitive	OBL = oblique	SBJ = subject
IPFV = imperfective	PFV = perfective	SG = singular
IRR = irrealis	PL = plural	SS = secondary suffix

References

- Ahland, Colleen. 2012. *A grammar of Northern and Southern Gumuz*. Eugene: University of Oregon. (Doctoral dissertation.)
- Ahland, Michael. 2012. *A grammar of Northern Mao (Màwés Aas'è)*. Eugene: University of Oregon. (Doctoral dissertation.)
- Amha, Azeb. 2001. *The Maale language*. Leiden: Universiteit Leiden. (Doctoral dissertation.)
- Anderson, Cormac, Natalia Kuznetsova & Shelece Easterday. 2025. Rarities in phonetics and phonology: Exploring rarities. In Natalia Kuznetsova, Cormac Anderson & Shelece Easterday (eds.), *Rarities in phonetics and phonology: Structural, typological, evolutionary, and social dimensions*, 3–28. Berlin: Language Science Press.
- Applebaum, Ayla Ayda Bozkurt. 2013. *Prosody and grammar in Kabardian*. Santa Barbara: University of California. (Doctoral dissertation.)
- Aronoff, Mark & Nanna Fuhrhop. 2002. Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint. *Natural Language and Linguistic Theory* 20(3). 451–490. DOI: 10.1023/A:1015858920912
- Asao, Yoshihiko. 2015. *Left-right asymmetries in words: A processing-based account*. Buffalo: State University of New York. (Doctoral dissertation.)
- Becker, Laura. 2024. Zero marking in inflection: A token-based approach. *Journal of Language Modeling* 12(2). 349–413.
- Belete, Alemgena. 2018. *Documentation and grammatical description of Kara*. Addis Ababa: University of Addis Ababa. (Doctoral dissertation.)
- Berdicevskis, Aleksandrs & Arturs Semenuks. 2020. Different trajectories of morphological overspecification and irregularity under imperfect language

- learning. In Peter Arkadiev & Francesco Gardani (eds.), *The complexities of morphology*, 283–305. Oxford: Oxford University Press.
- Berdicevskis, Aleksandrs & Arturs Semenuks. 2022. Imperfect language learning reduces morphological overspecification: Experimental evidence. *PLoS ONE* 17(1), e0262876. DOI: 10.1371/journal.pone.0262876
- Berg, Thomas. 2020. Ordering biases in cross-linguistic perspective: The interaction of serial order and structural level. *Linguistic Typology* 24(2). 353–397. DOI: 10.1515/Lingty-2019-2031.
- Berg, Thomas. 2022. Paradigmatic consequences of the suffixing preference. *Studies in Language* 46(2). 376–401. DOI: 10.1075/SI.20019.ber
- Bickel, Balthasar & Johanna Nichols. 2013. Inflectional synthesis of the verb. In Matthew Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <http://wals.info/chapter/22>. (Accessed May 14, 2025.)
- Bisang, Walter. 2020. Radical analyticity and radical pro-drop scenarios of diachronic change in east and mainland southeast Asia, west Africa and pidgins and creoles. *Asian Languages and Linguistics* 1(1). 34–70. DOI: 10.1075/Alal.00002
- Bisnath, Felicia, Hannah Lutzenberger, Marah Jaraisy, Rehana Omardeen & Adam Schembri. 2026. Deconstructing notions of morphological ‘complexity’: Lessons from creole and sign languages. *Journal of Linguistics* 62(1), 177–207. DOI: 10.1017/S0022226725000180
- Bobaljik, Jonathan. 2012. *Universals in comparative morphology: Suppletion, superlatives, and the structure of words*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/9069.001.0001
- Börstell, Carl. 2021. *GlottoshinyR*. Available at: <https://linguistr.shinyapps.io/glottoshinyr/>. (Accessed May 14, 2025.)
- Bybee, Joan. 2002. Main clauses are innovative, subordinate clauses are conservative. In Joan Bybee & Michael Noonan (eds.), *Complex sentences in grammar and discourse*, 1–17. Amsterdam: Benjamins.
- Bybee, Joan. 2006. Language change and universals. In Ricardo Mairal & Juana Gil (eds.), *Linguistic universals*, 179–194. Cambridge: Cambridge University Press.
- Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press. DOI: 10.1017/Cbo9780511750526

- Bybee, Joan. 2011. Markedness: Iconicity, economy, and frequency. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 131–147. Oxford: Oxford University Press.
- Bybee, Joan. 2015. *Language change*. Cambridge: Cambridge University Press. DOI: 10.1017/Cbo9781139096768
- Bybee, Joan, William Pagliuca & Revere Perkins. 1990. On the asymmetries in the affixation of grammatical material. In William Croft, Keith Denning & Suzanne Kemmer (eds.), *Studies in typology and diachrony: Papers Presented to Joseph H. Greenberg on his 75th Birthday*, 1–42. Amsterdam: Benjamins.
- Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: The University of Chicago Press.
- Colarusso, John. 1992. *A grammar of the Kabardian language*. Calgary: University of Calgary Press.
- Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2015. *The Leipzig glossing rules*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>. (Accessed September 18, 2025)
- Cristofaro, Sonia. 2014. Competing motivation models and diachrony: What evidence for what motivations? In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing motivations in grammar and usage*, 282–298. Oxford: Oxford University Press.
- Croft, William. 2003. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press. DOI: 10.1017/Cbo9780511840579
- Cysouw, Michael & Jan Wohlgemuth. 2010. The other end of universals: Theory and typology of rara. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking universals: How rarities affect linguistic theory*, 1–10. Berlin: De Gruyter.
- Dahl, Östen. 1979. Typology of sentence negation. *Linguistics* 17(1/2), 79–106.
- Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: Benjamins. DOI: 10.1075/Slcs.71
- Daniel, Michael & Yury Lander. 2011. The Caucasian languages. In Bernd Kortmann & Johan van der Auwera (eds.), *The languages and linguistics of Europe: A comprehensive guide*, 125–157. Berlin: De Gruyter Mouton.
- DeGraff, Michael. 2003. Against creole exceptionalism. *Language* 79(2). 391–410. DOI: 10.1353/Lan.2003.0114

- Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press. DOI: 10.1017/9781108671040
- Dixon, R. M. W. 2012. *Basic linguistic theory*. Vol. 3. Oxford: Oxford University Press.
- Dryer, Matthew. 2005. Negative morphemes. In Martin Haspelmath, Matthew Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 454–457. Oxford: Oxford University Press.
- Dryer, Matthew. 2013. Order of negative morpheme and verb. In Matthew Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <https://wals.info/feature/143>. (Accessed May 14, 2025.)
- Dryer, Matthew & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <http://wals.info> (Accessed September 14, 2025.)
- Enrique-Arias, Andrés. 2002. Accounting for the position of verbal agreement morphology with psycholinguistic and diachronic explanatory factors. *Studies in Language* 26(1). 1–31.
- Etxepare, Ricardo. 2003. Negation. In José Ignacio Hualde & Jon Ortiz De Orbina (eds.), *A grammar of Basque*, 516–564. Berlin: Mouton De Gruyter.
- Fenk-Oczlon, Gertraud. 1991. Frequenz und Kognition – Frequenz und Markiertheit. *Folia Linguistica* 25(3–4). 361–394.
- Frajzyngier, Zygmunt, Eric Johnston & Adrian Edwards. 2005. *A grammar of Mina*. Berlin: Mouton De Gruyter. DOI: 10.1515/9783110893908
- Gelderen, Elly van. 2008. Negative cycles. *Linguistic Typology* 12(2). 195–243. DOI: 10.1515/Lity.2008.037
- Givón, Talmy. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language* 15(2). 335–370. DOI: 10.1075/sl.15.2.05giv
- Givón, Talmy. 2001. *Syntax: An introduction*. Vol. 1. Amsterdam: Benjamins. DOI: 10.1075/Z.Syn1
- Greenberg, Joseph. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg (ed.), *Universals of language*, 2nd edn., 73–113. Cambridge, MA: The MIT Press.
- Greenberg, Joseph. 2005 [1966]. *Language universals, with special reference to feature hierarchies*. Berlin: Mouton De Gruyter.

- Griscom, Richard. 2019. *Topics in Asimjeeg Datooga verbal morphosyntax*. Eugene: University of Oregon. (Doctoral dissertation.)
- Grossman, Eitan & Stéphane Polis. 2018. Swimming against the typological tide or paddling along with language change? Dispreferred structures and diachronic biases in affix ordering. *Journal of Historical Linguistics* 8(3). 388–443. DOI: 10.1075/Jhl.17027.Gro
- Grossman, Eitan, Anton Antonov & Guillaume Jacques. 2018. A cross-linguistic rarity in synchrony and diachrony: Adverbial subordinator prefixes exist. *STUF* 71(4). 513–538.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2026. *Glottolog* 5.3. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <http://glottolog.org> (Accessed June 22, 2026.)
- Handschuh, Corinna. 2014. *A typology of marked-S languages*. Berlin: Language Science Press. DOI: 10.17169/Fudocs_Document_000000019858
- Harris, Alice. 2008. On the explanation of typologically unusual structures. In Jeff Good (ed.), *Linguistic universals and language change*, 54–76. Oxford: Oxford University Press.
- Harris, Alice. 2010. Explaining typologically unusual structures: The role of probability. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking universals: How rarities affect linguistic theory*, 91–103. Berlin: De Gruyter.
- Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42(1). 25–70. DOI: 10.1017/S0022226705003683
- Haspelmath, Martin. 2018. The last word on polysynthesis: A review article. *Linguistic Typology* 22(2). 307–326. DOI: 10.1515/Lingty-2018-0011
- Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633. DOI: 10.1017/S0022226720000535
- Himmelmann, Nikolaus. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960. DOI: 10.1353/Lan.2014.0105
- Hopper, Paul & Sandra Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–299.
- Jagersma, Abraham. 2010. *A descriptive grammar of Sumerian*. Leiden: Universiteit Leiden. (Doctoral dissertation.)
- Janhunen, Juha. 2012. *Mongolian*. Amsterdam: Benjamins. DOI: 10.1075/Loall.19

- Jing, Yingqi & Paul Widmer & Balthasar Bickel. 2023. Word order evolves at similar rates in main and subordinate clauses: Corpus-based evidence from Indo-European. *Diachronica* 40(4). 532–556. DOI: 10.1075/Dia.20035.Jin
- Kimball, Geoffrey. 1985. *A descriptive grammar of Koasati*. New Orleans: Tulane University. (Doctoral dissertation.)
- König, Christa. 2006. Marked nominative in Africa. *Studies in Language* 30(4). 655–732. DOI: 10.1075/Sl.30.4.02kon
- Kratochvil, František. 2007. *A grammar of Abui, a Papuan language of Alor*. Utrecht: LOT.
- Kumakhov, Muchadin. 1989. *Sravnitel'no-Istoricheskaja Grammatika Adygsikx (Cherkesskix) Jazykov*. [Comparative-historical grammar of the Adyghe (Cherkess) languages]. Moscow: Nauka.
- Kumakhov, Muchadin. 2006. *Kabardino-Cherkesskij Jazyk: V Dvukh Tomakh. Tom I*. [Kabardian-Cherkess language in two volumes. Volume 1: Writing system, phonetics and phonology, morphology, syntax]. Moscow: Nalchik.
- Kuteva, Tania, Bernd Heine, Bo Hong, Haiping Long, Heiko Narrog & Seongha Rhee. 2019. *World lexicon of grammaticalization*. 2nd edn. Cambridge: Cambridge University Press.
- Lacroix, René. 2009. *Description du dialecte Laze d'Arhavi (caucasique du sud, Turquie): Grammaire et textes*. [Description of the Laz dialect of Arhavi (South Caucasian, Turkey): Grammar and texts.] Lyon: Université Lumière Lyon 2. (Doctoral dissertation.)
- Lacroix, René. 2019. Laz. In Geoffrey Haig & Geoffrey Khan (eds.), *The languages and linguistics of western Asia: An areal perspective*, 830–860. Berlin: De Gruyter Mouton.
- Lehmann, Christian. 2004. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan, Stavros Skopeteas & Wolfgang Kesselheim (eds.), *Morphology: An international handbook on inflection and word-formation*, Vol. 2, 1834–1857. Berlin: De Gruyter.
- Lehmann, Christian. 2015. *Thoughts on grammaticalization*. 3rd edn. Berlin: Language Science Press. DOI: 10.5281/Zenodo.7520796
- Linn, Mary. 2001. *A grammar of Euchee (Yuchi)*. Lawrence: University Of Kansas. (Doctoral dissertation.)
- Malchukov, Andrej. 2011. Interaction of verbal categories: Resolution of infelicitous grammeme combinations. *Linguistics* 49(1). 229–282. DOI: 10.1515/Ling.2011.006

- Maslova, Elena. 2003. *A grammar of Kolyma Yukaghir*. Berlin: Mouton De Gruyter. DOI: 10.1515/9783110197174
- Mayerthaler, Willi. 1981. *Morphologische Natürlichkeit [Morphological Naturalness]*. Wiesbaden: Athenaion.
- Mayerthaler, Willi. 1987. System-independent morphological naturalness. In Wolfgang Dressler (ed.), *Leitmotifs in natural morphology*, 25–58. Amsterdam: Benjamins.
- McWhorter, John. 2005. *Defining creole*. Oxford: Oxford University Press.
- McWhorter, John. 2007. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford: Oxford University Press.
- Miestamo, Matti. 2005. *Standard negation*. Berlin: De Gruyter.
- Miestamo, Matti. 2007. Symmetric and asymmetric encoding of functional domains, with remarks on typological markedness. In Matti Miestamo & Bernhard Wälchli (eds.), *New challenges in typology: Broadening the horizons and redefining the foundations*, 293–314. Berlin: De Gruyter.
- Miestamo, Matti. 2010. Negatives without negators. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking universals: How rarities affect linguistic theory*, 169–194. Berlin: De Gruyter.
- Mihas, Elena & Gregorio Santos Pérez. 2015. *A grammar of Alto Perené (Arawak)*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110766301
- Mithun, Marianne. 2003. Why prefixes? *Acta Linguistica Hungarica* 50(1/2). 155–185.
- Ngo, Binh. 2021. *Vietnamese: An essential grammar*. Routledge: London.
- Nguyen, Đình Hòa. 1997. *Vietnamese*. Amsterdam: Benjamins. DOI: 10.1075/Loall.9
- Ozerov, Pavel. 2018. Tone assignment and grammatical tone in Anal (Tibeto-Burman). *Studies in Language* 42(3). 708–733.
- Ozerov, Pavel. 2019. Person indexation in Anal. *Himalayan Linguistics* 18(1). 26–53.
- Pilot-Raichoor, Christiane. 2010. The Dravidian zero negative: Conceptualisation and diachronic context of its morphogenesis. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rara and rarissima: Documenting the fringes of linguistic diversity*, 267–304. Berlin: De Gruyter.
- Shcherbakova, Olena, Susanne Maria Michaelis, Hannah Haynie, Sam Passmore, Volker Gast, Russell Gray, Simon Greenhill, Damián Blasi & Hedvig Skirgård. 2023. Societies of strangers do not speak less complex languages. *Science Advances* 9. DOI: 10.1126/Sciadv.Adf7704

- Skirgård, Hedvig, Hannah Haynie, Damián Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latache, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bower, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy Bakker, Giulia Barbos, Robert Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard Kowalik, Olga Krasnoukhova, Nora Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia Martins, Celia Mata German, Suzanne Van Der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney Rey, Julia Rizaew, Eloisa Ruppert, Kim Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário De Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye 葉婧婷, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen Levinson, Martin Haspelmath, Simon Greenhill, Quentin Atkinson & Russell Gray. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances* 9. DOI: 10.1126/Sciadv.Adg6
- Spencer, Andrew & Ana Luís. 2012. *Clitics: An introduction*. Cambridge: Cambridge University Press. DOI: 10.1017/Cbo9781139033763
- Thompson, Lawrence. 1987. *A Vietnamese reference grammar*. Honolulu: University of Hawai'i Press.
- Vázquez Álvarez, Juan Jesús. 2011. *A grammar of Chol, a Mayan language*. Austin: University of Texas. (Doctoral dissertation.)
- Velupillai, Viveka. 2012. *An introduction to linguistic typology*. Amsterdam: Benjamins.
- Weir, Helen. 1986. Diachronic development of certain verb prefixes in an OSV language (Nadëb). *Lingua* 68(4). 291–316.
- Willett, Thomas. 1991. *A reference grammar of Southeastern Tepehuan*. Dallas: Summer Institute of Linguistics.

- Wurzel, Wolfgang. 1987. System-dependent morphological naturalness in inflection. In Wolfgang Dressler (ed.), *Leitmotifs in Natural Morphology*, 59–96. Amsterdam: Benjamins.
- Wurzel, Wolfgang. 2001. *Flexionsmorphologie und Natürlichkeit: Ein Beitrag zur morphologischen Theoriebildung*. [Inflectional morphology and naturalness: A contribution to the formation of morphological theory]. 2nd edn. Berlin: Akademie Verlag. DOI: 10.1515/9783050080130
- Zeijlstra, Hedde. 2007. Negation in natural language: On the form and meaning of negative elements. *Language and Linguistics Compass* 1(5). 498–518. DOI: 10.1111/J.1749-818x.2007.00027.X
- Zingler, Tim. 2022. Clitics, anti-clitics, and weak words: Towards a typology of prosodic and syntagmatic dependence. *Language and Linguistics Compass*, e12453. DOI: 10.1111/lnc3.12453
- Zipf, George. 1965 [1935]. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The MIT Press.
- Zwicky, Arnold & Geoffrey Pullum. 1983. Cliticization vs. inflection: English *n't*. *Language* 59(3). 502–513. DOI: 10.2307/413900

Appendix. Languages surveyed.

The tables below list the languages for which I studied reference grammars. They are alphabetically sorted by macro-area and, within macro-areas, by language family. Glossonyms in square brackets reflect the names used by the relevant source in cases where that name differs from the one used in Glottolog. Boldfaced languages are the ones discussed in Section 3.

Family	Languages
Afro-Asiatic	Buwal, Coptic, Dangaleat [East Dangla], Ghomara (Berber), Goemai, Hdi, Jibbali, Konso, Mpade [Makary Kotoko], Mehri, Mina (Cameroon), Miya, Moloko, Sukur [Sakun], Senhaja de Srair [Senhaja Berber], Sidamo [Sidaama], Somali, Standard Arabic, Tamasheq [Tamashek], Tigre, Wandala
Atlantic-Congo	Aghem, Bena (Tanzania) [Bena], Bullom So [Mani], Chakali, Chuwabu [Cuwabo], Emai-Iuleha-Ora [Emai], Etulo, Ewe, Fon [Fongbe], Fwe, Ga [Gã], Gbeya, Gyele [Gyeli], Igbo, Ikpana (Logba), Liko, Lunda, Makonde, Mambai [Mambay], Mankanya, Mundabli-Mufu [Mundabli], Mungbam, Ncane-Mungong [Nchane], Ndengereko [Ndengeleko], Ndut, Noon, Southern Kisi [Kisi], Supyire Senoufo [Supyire], Swahili, Tafi, Tigon Mbembe [Mbembe]

Family	Languages
Bangime	Bangime
Blue Nile Mao	Bambassi [Northern Mao]
Central Sudanic	Ma'di
Dizoid	Nayi [Naayi], Sheko
Dogon	Jamsay Dogon [Jamsay], Tommo So Dogon [Tommo So]
Eastern Jebel	Gaam [Gaahmg]
Furan	Fur
Gumuz	Northern Gumuz, Southern Gumuz
Kadugli-Krongo	Krongo
Khoe-Kwadi	Kxoe [Khwe], Nama (Namibia) [Nama Hottentot], Northern Tshwa [Tjwao], Ts'ixa
Koman	Uduk
Kru	Grebo
Kuliak	Ik, Nyang'i
Kxa	Amkoe [ḥHōā]
Maban	Maba (Chad)
Mande	Jalkunan, Seeku [Seenku], Yalunka [Jalonke]
Narrow Talodi	Lumun
Nilotic	Isimjeega-Rootigaanga [Datooga (Asimjeeg)] , Lango (Uganda), Lopit, Luo (Kenya and Tanzania) [Dholuo], Luwo, Teso [Ateso], Turkana
Nubian	Kenuzi-Dongola [Kunuz]
Saharan	Dazaga, Kanuri (Central)
Sandawe	Sandawe
Shabo	Shabo [Chabu]
Songhay	Koyraboro Senni Songhai [Koyraboro (Koroboro) Senni], Koyra Chiini Songhay [Koyra Chiini], Tagdal
South Omotic	Dime, Hamar-Banna [Hamar], Karo (Ethiopia) [Kara]
Surmic	Majang, Mursi
Ta-Ne-Omotic	Kachama-Ganjule-Haro [Ganta], Male (Ethiopia) [Maale] , Zayse-Zergulla [Zaysete]

Table 6: Language sample: Africa. [28 families, 101 languages].

Family	Languages
Bunaban	Gooniyandi
Gaagudju	Gaagudju
Garrwan	Garrwa (Western)
Gunwinyguan	Bininj Kun-wok [Bininj Gun-wok], Anindilyakwa [Enindhilyakwa], Kunbarlang, Ngalakgan [Ngalakan], Rembarrnga [Rembarnga], Wubuy [Nunggubuyu]

Family	Languages
Iwaidjan Proper	Mawng
Mangarrayi-Maran	Mangarrayi
Maningrida	Guragone [Gurr-goni], Nakara [Nakkara]
Mirndi	Jingulu, Wambaya
Nyulnyulan	Bardi, Nyigina, Nyulnyul, Yawuru
Pama-Nyungan	Eastern Arrernte [Mparntwe], Dieri [Diyari], Djambarrpuyngu [Djamparrpuynu], Dyirbal, Gurindji, Guugu Yimidhirr, Jaru [Wanyjirra], Kuku-Yalanji, Mangala [Mangarla], Martuthunira, Ngardi, Ngarinman [Bilinarra], Ngiyambaa [Ngiyamba], Nhanda, Thayore [Kuuk Thaayorre], Martu Wangka [Wangkajunga], Warlmanpa, Warrongo, Wik-Mungkan, Wirangu, Yidiñ [Yidj], Yintyinka-Ayabadhu [Yintyingka]
Southern Daly	Murriny Patha, Nangikurrunggurr [Ngan'gityemerri]
Tangkic	Kayardild, Lardil [Yukulta]
Tiwi	Tiwi
Western Daly	Ami [Emmi], Marithiel [Marrihiyel]
Worrorran	Ngarinyin [Ungarinjin], Worrorra
Yangmanic	Wardaman

Table 7: Language sample: Australia. [16 families, 51 languages].

Family	Languages
Abkhaz-Adyge	Abkhaz, Kabardian
Austroasiatic	Chut [May], Ho, Jehai [Jahai], Kharia, Khmer (Central) [Cambodian], Khmu, Koho [Koho-Sre], Pacoh, Pnar, Semelai, Vietnamese
Basque	Basque
Burushaski	Burushaski (Eastern)
Chukotko-Kamchatkan	Chukchi
Dravidian	Kannada, Betta Kurumba, Kurukh [Kurux]
Great Andamanese	Akajeru
Hmong-Mien	Iu Mien, Western Xiangxi Miao [Xong]
Indo-European	Afrikaans, Bengali, Breton, Catalan, Czech, Danish, Domari, Dutch, Eastern Armenian, Eastern Panjabi [Panjabi], French, Greek, Hindi, Hittite, Icelandic, Irish, Laki, Latin, Latvian, Lithuanian, Maithili, Marathi, Northern Tosk Albanian, Norwegian, Polish, Portuguese, Rajbanshi, Romani, Romanian, Russian, Serbian-Croatian-Bosnian [Bosnian/Croatian/Montenegrin/Serbian], Sinhala, Slovenian [Slovene], Spanish, Swedish, Welsh, Western Balochi [Modern Standard], Western Farsi [Persian]
Japonic	Japanese, Miyako [Irabu]

Family	Languages
Jarawa-Onge	Jarawa (India) [Jarawa]
Kartvelian	Georgian, Laz
Koreanic	Korean, Jejueo
Kusunda	Kusunda
Mongolic-Khitani	Bonan [Bao'an Tu], Mangghuer, Halh Mongolian [Mongolian]
Nakh-Daghestanian	Hinuq, Hunzib, Ingush, Khinalug [Xinaliq], Khwarshi-Inkhoqwari [Khwarshi], Lezgian, North-Central Dargwa [Mehweb], Southwestern Dargwa [Sanzhi Dargwa]
Nivkh	Amur Nivkh
Sino-Tibetan	Ao Naga [Mongsen Ao], Bantawa, Burmese, Mandarin Chinese, Darma, Dhimal, Dumi, Dura, Ersu, Galo, Guqiong, Hills Karbi [Karbi], Japhug, Katso [Khatso], Kurtokha [Kurtöp], Lahu, Lepcha, Limbu, Manange, Manipuri [Meithei], Nocte Naga [Hakhun Tangsa], Purik-Sham-Nubra [Purik], Qiang, Sichuan Yi [Nuosu], Southern Pumi [Prinmi], Solu-Khumbu Sherpa [Sherpa], Sunwar, Tibetan, Tshangla, Western Parbate Kham [Kham], Yakkha
Sumerian	Sumerian
Tai-Kadai	Lakkia [Lakkja], Lao, Thai, Turung
Tungusic	Evenki, Manchu, Udihe
Turkic	Kazakh, Turkish
Uralic	Finnish, Hungarian, Mansi (Central), Nenets (Tundra), Nganasan, North Saami [North Sámi], Pite Saami, Skolt Saami, Vach-Vasjungan [Eastern Khanty]
Yeniseian	Ket, Yugh
Yukaghir	Northern Yukaghir, Southern Yukaghir [Yukaghir (Kolyma)]

Table 8: Language sample: Eurasia. [25 families, 134 languages].

Family	Languages
Algic	Arapaho, Cree (Plains), Cree (Woods), Ottawa [Nishnaabemwin], Malecite-Passamaquoddy [Passamaquoddy-Maliseet], Potawatomi, Shawnee, Siksika [Blackfoot], Yurok
Athabaskan-Eyak-Tlingit	Chilcotin-Nicola [Tsilhqút'in], Eel River Athabaskan [Wailaki], Hupa-Chilula [Hupa], North Slavey [Slave], Sarsi [Sarcee], Tanacross, Upper Tanana
Caddoan	Caddo
Chibchan	Pech, Rama, San Blas Kuna [Guna], Teribe
Chimariko	Chimariko
Cochimi-Yuman	Havasupai-Walapai-Yavapai [Hualapai], Maricopa, Tipai [Jamul Tiipay]
Eskimo-Aleut	Alaskan Iñupiaq (North), Aleut, Central Alaskan Yupik, Eastern Canadian Inuktitut, Kalaallisut [West Greenlandic], Seward Alaska Inupiatun [Iñupiaq]
Huavean	Huave (San Francisco del Mar), Huave (San Dionisio del Mar)

Family	Languages
Iroquoian	Cayuga, Cherokee (Oklahoma), Onondaga, Seneca, Huron-Wyandot [Wyandot, Wendat]
Karuk	Karuk [Karok]
Keresan	Western Keres [Acoma (Keresan), Laguna Keres]
Kiowa-Tanoan	Kiowa, Towa
Kutenai	Kutenai
Mayan	Chol, Itza' [Itzaj], Huastec (Southeastern, Potosino), Mam, Popti' [Jacaltepec], Sacapulteco [Sacapultec], Tz'utujil, Yucatec Maya
Misumalpan	Ulwa
Miwok-Costanoan	Mutsun
Mixe-Zoque	Chimalapa Zoque, Highland Popoluca [Sierra Popoluca], Northeast Chiapas Zoque, Tlahuitoltepec Mixe [Ayutla Mixe]
Molale	Molale [Molalla]
Muskogean	Alabama, Choctaw, Creek, Koasati, Mikasuki
Otomanguean	Atzingo Matlatzinca [Ocuilteco], Coatlán-Loxicha Zapotec, Nopala Chatino [Teotepec Eastern], Northern Pame, Sochiapam Chinantec [Sochiapan Chinantec], Tataltepec Chatino, Western Highland Chatino (Yaitepec), Zacatepec Chatino (Eastern), Zenzontepec Chatino
Pomoan	Kashaya, Southern Pomo
Sahaptian	Nez Perce, Northwest Sahaptin [Yakima Ichishkiin/Sahaptin]
Salishan	Bella Coola, Columbia-Wenatchi [Nxa'amxcin], Comox, Lillooet, Halkomelem [Musqueam], Spokane [Spokan], Squamish
Siouan	Assiniboine, Biloxi, Crow, Hidatsa, Lakota [Lakhota]
Tarascan	Purepecha
Totonacan	Filomena Mata Totonac, Huehuetla Tepehua, Tlachichilco Tepehua, Yecuatla [Misantla] Totonac
Tsimshian	Gitksan [Gitksan], Nisga'a [Nisgha], Southern-Coastal Tsimshian [Coast Tsimshian]
Uto-Aztecan	Comanche, Cupeño, Huarijio [River Warihío], Hopi, Northern Paiute, Panamint [Tümpisa Shoshone], Pipil, Tarahumara (Central) [Choguita Rarámuri], Southeastern Tepehuan, Tohono O'odham, Ute-Southern Paiute [Ute]
Wakashan	Nuu-chah-nulth
Washo	Washo
Wintuan	Wintu
Yuchi	Yuchi
Yuki-Wappo	Wappo, Northern Yukian [Yuki]

Table 9: Language sample: North America. [33 families, 113 languages].

Family	Languages
Abun	Abun
Angan	Menya
Anim	Marind [Coastal Marind]
Austronesian	Acehnese, Balantak, Belep, Bunun, Cebuano, Chamorro, Dakaka [Daakaka], Dela-Oenale [Dela], East Makian [Taba], Fijian (Boumaa), Filipino, Hawaiian, Irarutu, Kambera, Kéo, Kilivila, Kokota, Kosraean [Kusaiean], Lamaholot, Madurese, Makasar [Makassarese], Malay (Papuan), Malay (Baba), Mafea [Mavea], Manam, Maori, Muna, Nafsan [South Efate], Neverver, Paiwan, Papanana, Pohnpeian [Ponapean], Puyuma, Rapanui [Rapa Nui], Saaroa [Lha'alua], Standard Indonesian [Indonesian], Tagalog, To'abaita [Toqabaqita], Tsat [Hainan Cham], Tukang Besi, Tuvalu [Tuvaluan], Ughele, Vaeakau-Taumako, Vurës, Western Subanon
Baibai-Fas	Momu-Fas [Momu]
Baining	Mali, Qaqet
Bilua	Bilua
Border	Imonda, Kilmeri
Bosavi	Aimele [Eimela]
Bulaka River	Wanam Yelmek
Duna	Duna
East Bird's Head	Meyah, Moskona
East Strickland	Konai
Fuyug	Fuyug
Hatam-Mansim	Hatam
Inanwatan	Suabo
Keram	Pondi, Ulwa (Papua New Guinea)
Kuot	Kuot
Lakes Plain	Abawiri
Lavukaleve	Lavukaleve
Lower Sepik	Kopar, Yimas
Manubaran	Doromu-Koki
Maybratic	Maybrat
Ndu	Ambulas, Iatmul, Manambu
Nimboranic	Nimboran
North Bougainville	Rotokas
North Halmahera	Ternate, Tidore
Nuclear Torricelli	Abu' Arapesh [Abu'], Bukiyip, Kamasau, Yeri

Family	Languages
Nuclear Trans New Guinea	East Kewa [Kewapi], Fore, Huli, Kesawai [Koromu (Kesawai)], Lower Grand Valley Dani, Ma Manda, Mauwake, Mian, Oksapmin, Tauya, Usan, Wano, Yagaria [Hua], Yau-Nungon
Savosavo	Savosavo
Senagi	Dera (Indonesia) [Menggwa Dla]
Sepik	Alamblak, Awtuw, Mehek
Sko	Bauni [Barupu], I'saka, Skou, Wutung
South Bougainville	Siwai [Motuna]
Taulil-Butam	Taulil [Tulil]
Taiap	Taiap [Tayap]
Timor-Alor-Pantar	Abui, Adang, Bunak [Bunaq], Klon, Makasae-Makalero, Reta, Teiwa
West Bomberai	Kalamang
Yam	Komnzo, Ngkontar Ngkolmpu [Ngkolmpu]
Yele	Yele

Table 10: Language sample: Papunesia. [40 families, 120 languages].

Family	Languages
Araucanian	Mapudungun [Mapuche]
Arawakan	Achagua, Apurinã, Ajiyíninka Apurucayali [Axininca Campa], Ashéninka Perené [Alto Perené (Arawak)], Baure, Garifuna, Lokono [Arawak Lokono Dian], Nanti, Parecís [Paresi-Haliti], Paunaka, South Ucayali Ashéninka [Ashéninka], Tariana, Yine (Piro)
Arawan	Culina [Kulina], Madi [Jarawara]
Aymaran	Southern Aymara [Aymara Muylaq'], Jaqaru
Barbacoan	Awa-Cuaiquer [Awa Pit]
Boran	Bora [Miraña]
Bororoan	Bororo
Cahuapanan	Jebero
Camsá	Camsá [Kamsá]
Cariban	Bakairí [Bakairi], Galibi Carib [Carib], Hixkaryána [Hixkaryana], Panare, Trió [Tiriyó], Waimiri-Atroari, Wayana
Chapacuran	Wari'
Chicham	Aguaruna [Aguaruna (Iiniá Chicham)], Achuar-Shiwiari [Shiwiari], Huambisa [Wampis]
Guaicuruan	Kadiwéu, Mocoví, Pilagá, Toba
Hibito-Cholón	Cholón
Huitotoan	Murui Huitoto [Murui Bue]

Family	Languages
Kakua-Nukak	Kakua
Kwaza	Kwaza
Lengua-Mascoy	Enxet Sur
Máku	Máku
Mataguayan	Nivaclé [Nivacle], Vezoz [Wichí], Weenhayek [Wichi]
Mosetén-Chimané	Mosetén-Chimané [Mosetén]
Movima	Movima
Naduhup	Hup
Nambiquaran	Mamaindé [Mamaindê], Sabanê
Nuclear-Macro-Je	Apinayé [Apinajé], Karajá, Rikbaktsa
Pano-Tacanan	Cashibo-Cacataibo [Kakataibo], Cavineña, Chácobo, Ese Ejja, Matsés [Matses], Resígaro, Shipibo-Conibo [Shipibo-Konibo]
Peba-Yagua	Yagua
Puinave	Puinave
Quechuan	Huallaga Huánuco Quechua, Yauyos Quechua
Saliban	Maco [Mako]
Ticuna-Yuri	Tikuna [Ticuna]
Trumai	Trumai
Tucanoan	Cubeo [Kubeo], Desano, Kotiria
Tupian	Akuntsu [Akuntsú], Cocama-Cocamilla [Kukama-Kukamiria], Kamayurá [Kamaiurá], Karitiâna [Karitiana], Karo (Brazil), Mundurukú, Paraguayan Guaraní [Guaraní], Tapieté [Tapiete], Tuparí
Urarina	Urarina
Yanomamic	Ninam [Yanam (Shiriana)], Yanomám
Yuracaré	Yuracaré [Yurakaré]

Table 11: Language sample: South America. [37 families, 85 languages].

CONTACT

tim.zingler@uibk.ac.at

Habitual markers and their lexical sources across languages

SUNE GREGERSEN¹, EVA VAN LIER²

¹UPPSALA UNIVERSITY, ²UNIVERSITY OF AMSTERDAM

Submitted: 03/09/2024 Revised version: 30/05/2025

Accepted: 26/06/2025 Published: 23/07/2026



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

Abstract

This paper presents a cross-linguistic study of habitual markers, defined as morphs or constructions expressing that a situation typically occurs. We investigate the frequency, formal types, and lexical origins of dedicated habitual markers in the world's languages. On the basis of an areally and genealogically balanced sample of 102 languages, we find that such markers are cross-linguistically common, being reported in 52 of the languages in the sample; in 12 of these, more than one habitual marker was found. Affixes are the most commonly reported type of marker, followed by particles and auxiliaries. For the investigation of the lexical origins of habitual markers, we added a convenience sample of 85 languages and dialects where we knew habituais to be attested. This allowed us to collect a large catalogue of lexical sources of such markers, which includes lexemes meaning 'know', 'do', 'like/love', 'belong', 'be strong', and various verbs of movement.

Keywords: habituality; aspect; grammaticalization; semantic change; typology.

1. Introduction

Habitual markers are aspectual expressions indicating that a situation typically or usually occurs. In this paper, we present the results of a cross-linguistic survey of such markers, focussing in particular on the formal expression and likely diachronic origins

of habitual morphs and constructions. To this end, we surveyed descriptions of a variety sample of 102 genealogically and areally diverse languages, to which we added a convenience sample of 85 languages for which we knew information on habitual markers to be available. The findings from the variety sample suggest that dedicated markers of habitual aspect are quite frequent in the languages of the world, being attested in 52 of the 102 languages. The two samples combined allow us to create a catalogue of a large number of lexical items which may develop into habitual markers; in this way, our investigation supplements the lists of lexical sources (“grammaticalization pathways”) collected by Bybee et al. (1994) and Kuteva et al. (2019).

The paper is structured as follows. Section 2 gives a brief introduction to habitual meaning and some of the central literature on this semantic category. In Section 3 we explain the goals, methods, and language samples used for the investigation. Sections 4 and 5 then present the findings of the study, the former focussing on the frequency and formal characteristics of habitual markers, the latter on the likely diachronic sources of these markers. The discussion in Section 6 places our findings in the context of the earlier literature and points out some directions for future research on habitual markers. Section 7 is a brief conclusion.

2. On habituals

Habitual markers have been described for numerous languages of the world, and the term ‘habitual’ itself is commonly used in both cross-linguistic and language-specific work on aspect (e.g. Comrie 1976; Dahl 1985; Nurse 2008; Romeo 2008; Mattiola 2019, 2020; and many others). While there is no universally agreed upon definition of habitual meaning (or ‘habituality’), most scholars would probably agree that e.g. English *used to*, Spanish *soler* and the Yuchi affix *-ne* in (1a) count as examples of habitual markers. In Yuchi, the suffix *-ne* is used to express that a state of affairs “takes place regularly or frequently over a long period of time” (Linn 2001: 260), whereas verbs expressing episodic states of affairs, as in (1b), do not take this suffix:

(1) Yuchi (YUC; isolate, USA; Linn 2001: 263, 77)

- a. *sahã:de* *’i* *hẽ-wa-ne*
 always tobacco 3SG(YM).ACT-chew-HAB
 ‘He chews tobacco all the time’

- b. *ʔ* *hõ-wa = :le*
 tobacco 3SG(YM).ACT-chew = EMPH
 ‘He’s really chewing (and chewing) that tobacco!’

Habituality has often been described as a subtype of imperfective or pluractional meaning. Indeed, Bertinetto & Lenci (2012: 854) place habituality “at the intersection of pluractionality and gnomic imperfectivity”, the latter term covering expressions of ‘law-like’ generalizations about the world. Similar characterizations are found in several other works on verbal aspect, such as Bybee et al. (1994: 124), Xrakovskij (1997a: 53), and Mattioli (2019: 31–32), who discuss habitual meaning from different angles. The famous study by Bybee et al. (1994) uses a 76-language sample to investigate a number of TMA categories and their possible diachronic sources. Concerning habituals, the authors find 26 languages with one or more overt habitual markers, of which the marker in 9 languages is restricted to the past tense. The most commonly attested lexical sources are verbs meaning ‘know’ and ‘live’, but in most cases it is not possible to determine the origin of the marker in question. In the edited volume by Xrakovskij (1997b), habituality is considered as one of several types of ‘plurality of situations’, specifically one involving “some empirically observed probabilistic regularity” (Xrakovskij 1997a: 58). In a similar way, in Mattioli’s (2019) study of pluractional markers, habituality is described as a kind of ‘non-prototypical’ pluractionality which is closely related to, but distinct from, frequentativity, the repetition of the same state of affairs on different occasions. In some languages, the same marker may have both frequentative and habitual readings, whereas in others these meanings are distinguished formally. For Plains Cree, for instance, Wolvengrey (2026) describes a ‘frequentative–multiplicative’ construction with the formative *(C)āh~* (2a), which is distinct from the habitual construction with the particle *māna* (2b):

(2) Plains Cree (CRK; Algic, Cree; Wolvengrey 2026: 97, 107)

- a. ... *ē-kī-āh~it-iko-cik* *ēsa*
 CNJ-PST-FREQ~say.to-INV-3PL RPT
 ‘... they had repeatedly been told by them ...’
- b. ... *kī-itwē-wak* *ēsa* *māna* *kēhtē-ay-ak*
 PST-say-3PL RPT HAB old-one-PL
 ‘... the old people would say.’

Similarly, as Bybee et al. (1994: 125) note, habitual meaning “is sometimes expressed by a more general imperfective or present gram and sometimes has its own expression”. The present study focusses on the latter type, i.e. languages with markers dedicated to the expression of habitual meaning.

While detailed studies of habituality are available for a number of individual languages (see e.g. the contributions to Boneh & Jędrzejowski 2019a; Gregersen & Hengeveld 2026), there seems to be little cross-linguistic work devoted specifically to habitual markers. The few that we are aware of focus on very particular issues or rely on less than ideal language samples. Cristofaro (2004) focusses on the coexpression of habitual aspect and “various types of unactualized situation[s]” (Cristofaro 2004: 257), such as future, modalized, or negated events, which are subsumed under the label ‘irrealis’. This kind of coexpression was found in less than twenty languages in “a convenience sample of about two hundred languages” (Cristofaro 2004: 261). The relevant languages are listed in Cristofaro’s (2004: 260, 262) tables 1 and 2. Unfortunately, no information is given on the entire 200-language sample, and sources are only provided for a subset of the languages mentioned in the paper, making it difficult to verify Cristofaro’s general conclusions or carry out a follow-up study.¹ A recent study by Fortuin (2023) focusses on the relation between habituais and the perfective/imperfective distinction, arguing – against the received opinion – that habitual aspect should not be considered a subtype of imperfectivity.² The language sampling in this paper is much more transparent, but not areally and genealogically balanced, as the author acknowledges (Fortuin 2023: 12). Of the 36 languages in the sample, more than half are spoken in Eurasia (as defined by Glottolog), and more than a third are Indo-European. Finally, Carlson (2012) provides a survey of habitual expressions in various languages, but this is not based on any principled sample, and no overview is given of the languages surveyed by the author.

As mentioned above, habituais are discussed in some studies of tense and aspect categories, but have often played a somewhat marginal role. For instance, in Dahl’s (1985) questionnaire-based investigation of tense and aspect systems in 64 languages, it is concluded that few languages have habituality as a “major TMA category” (on

¹ A potentially interesting point is that among the 17 languages with habitual–irrealis coexpression mentioned by Cristofaro, only a single one (West Greenlandic; *KAL*; Eskimo-Aleut, Inuit) is native to the Americas. Because no information is provided on the entire sample, it is not possible to check whether this is accidental or might possibly reflect an areal trend.

² Though see also Dahl (1985: 78–79) and Boneh & Jędrzejowski (2019: 5–7) for similar views.

the definition of this term see Dahl 1985: 52–53), and that the questionnaire returned “too few instances of HAB [= habituals] to make clear statements about marking types” (Dahl 1985: 96). On a similar note, Thieroff (2000: 295–297) concludes that “fully grammaticalized” habitual markers are rare in the languages of Europe, as they are only attested in six of the forty languages in the sample. There are no chapters on habitual markers in either WALS (Dryer & Haspelmath 2013) or Grambank (Grambank Consortium 2023). The *World Lexicon of Grammaticalization* (Kuteva et al. 2019) catalogues a number of diachronic sources of habitual markers, such as verbs meaning KNOW, REMAIN, USE, and GO. For several of the pathways discussed, however, it is noted that the historical development is still poorly understood, e.g. the development of aspectual markers from existentials or the development from USE to a habitual marker (Kuteva et al. 2019: 165, 450).

3. Goals and methods

Because of the apparent lack of a cross-linguistic survey dedicated to habitual markers, we decided to carry out the investigation presented here. Our survey had two main goals. First, we wanted to investigate the frequency, distribution, and expression formats of dedicated habitual markers in the world’s languages, in other words to survey their diversity across the world. Second, we wanted to gather a catalogue of their likely lexical sources, thereby supplementing earlier work on grammaticalization by Bybee et al. (1994) and Kuteva et al. (2019).

For the first goal, we surveyed a variety sample consisting of 102 languages from different language families (see Appendix 1). The basis for this sample was composed by means of the Diversity Value Technique developed by Rijkhoff and Bakker (Rijkhoff & Bakker 1998; Bakker 2011) for the purpose of a recent study of verbal diminutives (Audring et al. 2021). Diversity Values, computed on the basis of nodes in language family tree, reflect a family’s internal complexity, thereby determining how many languages of that family should be included in a sample of a given size. However, this is irrelevant for a 102-language sample, in the sense that it constitutes the minimum one-language-per-family sample. This also clarifies why the sample is based on the Ethnologue classification (Lewis et al. 2015), rather than the more generally accepted Glottolog one (Hammarström et al. 2024): since the number of families in the latter is much larger than 102, a minimal sample would be much larger too. For practical reasons, a larger sample is not always feasible, and this applies to

our study too. Notably, however, we modified the sample from Audring et al. (2021) in a few ways. First, some languages were exchanged in cases where we managed to find more recent or more comprehensive sources for other languages from the same language family. Second, while the original sample was based on Ethnologue, in cases of disagreement between Ethnologue and Glottolog, we have instead followed the latter classification. This prompted us to remove two languages (Mbugu and Palula) whose families according to Glottolog were already represented by other languages in the sample.³ On the other hand, following Glottolog allowed us to include three additional languages from Australia, a macroarea which is underrepresented if one relies solely on Ethnologue.⁴

The information gathered from the variety sample was also useful for our second goal, i.e. the investigation of the lexical sources of habitual markers. However, dedicated habitual markers were not found in all languages in the sample, and for many of the attested markers we are not able to say anything about their possible diachronic origin. For this reason, we added a convenience sample of 85 additional languages and dialects for which we knew information on one or more habitual markers to be available (see Appendix 2). We encountered some of these while surveying the literature on habituals, whereas others were pointed out to us by other linguists during our survey. We have deliberately included many languages with periphrastic habitual constructions (e.g. auxiliaries and serial verbs) in the convenience sample, as the diachronic origins of such markers can be expected to be more transparent than e.g. affixes and clitics. Hence, the convenience sample is not representative of the full diversity of habitual strategies in the world's languages, and we will not make any quantitative generalizations on the basis of it.

For each language in the two samples, we surveyed relevant descriptive works and noted the following information relating to habitual markers:

1. Whether one or more dedicated habitual markers are described for the language.
2. The expression format of each habitual marker.
3. The likely diachronic source of each habitual marker.

³ Mbugu: MHD; Atlantic-Congo, Benue-Congo. Palula: PHL; Indo-European, Indo-Iranian.

⁴ In other words, while all 102 languages in the variety sample belong to different families according to the Glottolog classification, a few of them belong to the same families according to Ethnologue. For the sake of transparency we have included the genealogical information from both of these sources in the appendices. The genealogical information in the paper is taken from Glottolog (Hammarström et al. 2024).

4. Other relevant information mentioned in the source, e.g. tense or co-occurrence restrictions, semantic differences between markers, etc.

Although ‘habitual’ is a commonly known and used linguistic term, not all linguists use it in the same way, and the markers labelled ‘habitual’ in different languages are not necessarily functionally equivalent. In a cross-linguistic survey of this kind, the identification of the relevant phenomena of course depends on the analysis presented in the grammars and other sources used. However, we have done our best to check that the use of the marker in the linguistic examples corresponds to our definition of ‘habitual’. Our definition is given in (3):

(3) Habitual marker

A morph or construction expressing that a situation typically occurs.

This definition excludes a number of phenomena. As already mentioned above, we have limited ourselves to what we refer to as dedicated habitual markers, i.e. forms dedicated to the expression of habitual meaning, rather than imperfective or pluractional forms which may occur in habitual contexts.⁵ In other words, we have investigated habituality as a structural phenomenon rather than a substance category. This enables us to survey the possible lexical sources of habitual markers specifically, rather than e.g. imperfective or pluractional markers. Hence, aspect markers with a more general imperfective meaning were not included, such as the Semelai imperfective reduplication described by Kruspe (2004: 112–114); while this may occur in habitual contexts like (4a), it is also used for progressive situations, e.g. (4b).

(4) Semelai (SZA; Austroasiatic, Aslian; Kruspe 2004: 310, 113)

- a. *gn~gɔŋ* *ɲɔʔ*
 IPFV~bite always
 ‘(It) always bites.’

⁵ As independently pointed out to us by Andrey Shluinsky and an anonymous reviewer, it is also theoretically possible that habituality is a structural category in a language, but expressed through the absence of any marker. We have not come across any such ‘zero habituals’ in our survey. Note that the English simple present, which may be used in habitual contexts, also expresses a number of other aspectual values (cf. e.g. Quirk et al. 1985: 179–183; Carlson 2012: 832–833) and hence would not count as a dedicated habitual construction in our terms.

- b. *kəhn* *pk~pək*
 3SG IPFV~wash.clothes
 ‘She is washing clothes.’

Similarly, we did not include pluractional markers in the survey, which in some languages may also occur in habitual contexts (cf. Xrakovskij 1997b; Mattiola 2019). One example of this is CVC reduplication in Squamish (Skwxwú7mesh; SQU; Salishan, Coast Salish) as described by Bar-el (2008). While this construction is sometimes translated ‘usually, always’ (Bar-el gives the example *kw’elh~kw’elh* ‘always spill’), this may be regarded as an instance of its more general pluractional meaning ‘do X more than once’, as found also in *7exw~7exw* ‘cough many times’ (iterative) and *sek’~sak’* ‘slice’ (event-internal plurality, from *sak’* ‘cut’). For the purposes of this investigation, Squamish CVC reduplication would thus not be considered a habitual marker. In Section 6, we briefly discuss some of the possible differences between dedicated habitual and pluractional markers.

Adhering to our definition in (3), we have occasionally decided to exclude markers or constructions labelled ‘habitual’ in one of our sources. For instance, we have not included the Chukchi (CKT; Chukotko-Kamchatkan) ‘habitual’ described by Dunn (1999: 194–195), which appears to be a more general imperfective marker, or the Doromu (KQC; Manubaran) ‘habitual’ which according to Bradshaw (2012: 85) “denotes an on-going, continuous event” in the past. Similarly, “habitual nouns” (Wang et al. 2006: 32) or agent nominalizers were not included. On the other hand, markers have sometimes been included which are not described with the term ‘habitual’ in the source. While this term is clearly the most common one in the literature, several other terms are also encountered, especially but not exclusively in some older works. Some examples of alternative terminology – some from languages included in the survey, others from other sources which we have consulted – are given in Table 1. We discuss some of these works in more detail in Section 6.

In the majority of cases, we surveyed the most comprehensive grammar of a given language, focussing on chapters covering aspectual and related categories, verb morphology, and auxiliary constructions. While we have done our best to include all (and only the) relevant markers in our survey, in some instances one could undoubtedly have made different analytical choices. For a few languages, we managed to find studies dedicated to habitual expressions (or the aspectual system more broadly), which we of course consulted as well. In several cases we have relied on information from experts working on

particular languages. Finally, in the case of Japanese (JPN; Japonic), where the picture presented by the linguistic literature was not entirely clear, we conducted a small elicitation task with three native speakers in order to gather additional information.⁶

Term	Source(s)
CONSTANT FEATURE	Smeets 2008
CONSUEUDINAL	O'Donovan 1845; Bliss 1972
CUSTOMARY	Kennard 1936; Lipkind 1945; Hancock 1971; Young 1971; Miyaoka 2012; Mihas 2015; Ennever 2021
HABITIVE	Janhunen 2012
HABITUATIVE	de Angulo & Freeland 1930; Strange 1973; Louwerse 1988; Hill 1975
NORMAL	Bromley 1981; Louwerse 1988
TRADITIONAL	Jacobsen 1964
USITATIVE	Kennard 1936; Jacobsen 1964; Ultan 1967; Parks 1976; Silverstein 1974; Klokeid 1976; Nedjalkov & Otaina 2013

Table 1: Alternative terms for habitual markers.

4. Habitual markers across the world

In this section we present the findings on the frequency and expression format of habitual markers in the 102-language variery sample, i.e. the first two research questions mentioned in Section 3.

4.1. Frequency and distribution of habitual markers

The first question is whether one or more habitual markers are attested in the languages in the sample. This question could be answered positively for 52 out of the 102 languages, i.e. about half of them. Out of these 52 languages, we found reports

⁶ The elicitation was carried out in Osaka between 11 and 18 June 2024. The participants were three BA students of Danish (DAN; Indo-European, Germanic) at the University of Osaka. The elicitation task consisted of translation and discussion of a number of written English sentences with habitual contexts; additional clarification and explanation was provided in Danish. See the discussion of (18) below for the main conclusion.

of more than one habitual marker in 12 languages.⁷ The distribution of these according to the six macroareas is shown in Table 2.

Macroarea	Languages in the variety sample	Languages with at least one habitual	Languages with more than one habitual
South America	33	22	5
North America	25	13	2
Eurasia	19	9	3
Papunesia	15	5	1
Africa	6	2	–
Australia	4	1	1
Total	102	52	12

Table 2: Habitual markers per macroarea in the variety sample.

Note that we have only included languages in Table 2 in which we found positive evidence of at least one habitual marker according to our definition in (3). In a few cases, the available source merely mentioned the existence of a habitual marker without providing any further analysis or examples; see e.g. Briley (1997: 129) on Bauzi (BVZ; Geelvink Bay). These have not been included in the count in Table 2. In other cases, we found no mention of any habitual markers, but this of course does not rule out the possibility that the language may have one. In particular in cases where the only available source is a sketch grammar, it is quite possible that a marker has been overlooked.

4.2. Formal characteristics of habitual markers

For each of the habitual markers identified in the survey, we noted the expression format, i.e. the formal type of marker. Because 12 languages in the sample were found to have more than one habitual marker, the total number of markers ($n=75$) is greater than the number of languages with habituals ($n=52$). In Table 3 we list the number

⁷ In the context of our survey, “more than one habitual” means that there is more than one form expressing habitual meaning and that these forms are not (synchronically) derived from each other. In some cases, the habituals may be used more or less interchangeably; in other cases they form a paradigm, e.g. positive vs. negative or past vs. non-past habitual. For further discussion see Section 6.

of markers in the sample belonging to the different types of expression format; each type will be exemplified below the table.⁸

Type of marker	SAm	NAm	Euras	Papu	Afr	Aus	Total
Affix	14	14	4	3	–	2	37
Auxiliary	9	–	–	1	–	–	10
Particle	1	1	5	–	2	–	9
Clitic	4	1	–	–	–	–	5
Serial verb construction	2	–	1	1	–	–	4
Complex construction	2	–	2	–	–	–	4
Reduplication	1	–	–	–	–	–	1
Variable/unclear	2	–	2	1	–	–	5
Total	35	16	14	6	2	2	75

Table 3: Expression format of habitual markers in the variety sample.

As Table 2 shows, **affixes** are the most commonly reported type of habitual marker, accounting for about half of the markers in the sample (37 out of 75). An example from Yuchi was given in (1a) in Section 1. Another example, the Yagua habitual suffix *-sara*, is shown in (5):

(5) Yagua (YAD; Peba-Yagua; Payne & Payne 1990: 392)

ramyééchara

ray-mááy-**sara**

1SG-sleep-HAB

‘I’m always sleeping.’

There were ten examples of habitual **auxiliaries** in five languages in the sample. An example is Cófán/A’ingae *atesû*, which means ‘know (how to)’ as an independent verb, but is also used as a habitual auxiliary in the construction illustrated in (6):

⁸ We rely on our sources for the classification of expression formats, but of course recognize that the boundaries between these are not always clear-cut, in particular in the case of particles vs. clitics and auxiliaries vs. serial verbs (the latter case is discussed in Gregersen & van Lier *forthc.*). A combined “particle-clitic” category would result in 14 examples in the material, as would a combined “multi-verb” category consisting of auxiliaries and serial verb constructions.

- (6) Cofán/A'ingae (CON; isolate, Colombia/Ecuador; Hengeveld & Fischer 2018: 335)
Je'nda pûshesû = ndekhû = khe = ti setha'pue-ñe atesû = 'fa.
 then woman = HUMPL = ADD = Q sing-INF HAB = PLS
 'So the women too used to sing?'

Habitual **particles** were attested nine times in seven different languages. For instance, Kiowa uses a dedicated particle *àn* to express habitual meaning:

- (7) Kiowa (KIO; Kiowa-Tanoan; Watkins 1984: 192)
hò· jó-hò· àn è-kíl
 yes there-DEF HAB 1DU/PL-dwell
 'Yes, that's where we used to live.'

Habitual **clitics**, such as the Nuu-chah-nulth enclitic = *?a:la* in (8), were found five times in four different languages, all of them spoken in the Americas.

- (8) Nuu-chah-nulth (NUK; Wakashan, Southern Wakashan; Davidson 2002: 202)
má?a?ama?a:la
má.-'a?a = ma' = ?a:la
 cold-on.rocks = IND = HAB
 'It is always cold on the rocks.'

Habitual **serial verb constructions** were found in four languages. An example is the Lao verb *mak1* 'like', which may be used in a serial verb construction with habitual meaning (glossed 'tend' or 'TEND' by Enfield 2007):

- (9) Lao (LAO; Tai-Kadai, Daic; Enfield 2007: 175)
thèèw3 nii4, khaw5 bòø mak1 ngaam2.
 area DEM rice NEG tend look.good
 'In this area, rice tends not to be of good quality [lit. 'look good'].'

In three languages, we found other **complex constructions** (four in total), which we have grouped under this label for lack of a better term. They all involve some kind of biclausal structure, such as the Japanese past habitual construction exemplified in

(10). This uses the sentential nominalizer *mono* (< ‘thing’), which nominalizes the preceding clause, followed by a copula:

(10) Japanese (Japonic; Horie 2008: 176)

[*Yoku mukasi=wa umi=ni oyogi=ni it-ta*] *mono=da*.

often past=TOP sea=DAT swim.CONJ=DAT go-PST NMLZ=COP

‘In the past I would often go swimming in the sea’ (lit. ‘It’s that in the past I often went to the sea for swimming’)

The rarest individual type of marker in the sample involves **reduplication**, which was only found in the South American language Kwaza. Note that the construction in question is in fact more complex than prototypical reduplication, as it involves the reduplication of the person index on the predicate, the addition of a dependent-clause suffix, and the cliticization of a person marker and mood suffix to this dependent clause, as in (11):

(11) Kwaza (XWA; isolate, Brazil; van der Voort 2026: 188)

itso-na ũi-xa~xa-ta=xa-re

hammock-LOC lie-2SG~2SG-DEP=2SG-Q

‘Do you always sleep in a hammock?’

The construction (on which see van der Voort 2003, 2004: 455–456, 2026) is similar to the Japanese past habitual construction in (10) in that the habitual predicate is embedded in a dependent clause, but because it also involves reduplication, we have listed it under this heading in Table 3. We discuss the apparent rarity of reduplication in Section 6.

Finally, we have listed five strategies in the material (in five different languages) as variable/unclear because they could not be assigned to one of the other formal types, either because the grammar provided too little information or because the strategy in question shows formal variation in the language. An example of the latter case is the habitual marker *bíg ~ -bi* in Hup, which according to Epps (2008: 144) represents “two marginally distinct forms of the same morpheme”, namely a particle and a reduced suffix form; compare (12a) and (12b):

(12) Hup (JUP; Naduhup, Eastern Naduhup; Epps 2008: 145–146)

- a. *ʔãh hám-áy bíg*
 1SG go-DYN HAB
 ‘I go regularly’
- b. *ʔãh ham-bí-h*
 1SG go-HAB-DECL
 ‘I go regularly’

We thus find a variety of ways to express habitual meaning in the 102-language sample, with affixes being the most commonly attested strategy (on which see Section 6). The high frequency of affixes in the variety sample turned out to be somewhat problematic for the other main objective of the survey, namely to identify the likely diachronic sources of habitual markers. All else being equal, affixes can be expected to have higher time depths and less transparent diachronic sources than e.g. auxiliaries or serial verbs. Because most of the languages in the variety sample do not have a historical record, it is impossible to trace the history of the markers in question directly, and for most of the affixes in the sample we have no information about their diachronic origin. By contrast, for all auxiliaries and serial verbs in the sample, the source lexeme was identified in the source; see the examples from Cofán/A’ingae and Lao in (6) and (9). For this reason, we decided to survey an additional 85 languages, as explained in Section 3. The findings from the combined samples will be presented in the following section.

5. Lexical sources

5.1. Overview

There is relatively little work on the diachronic development of habitual markers, and for languages without a written record such work will necessarily have to rely on internal reconstruction or comparison with related languages. For several of the lexical sources of habitual markers listed by Kuteva et al. (2019), the authors add remarks to the effect that “more research is needed on this pathway” (Kuteva et al. 2019: 262; cf. also pp. 245, 347, 450). We agree that more diachronic work on habitual constructions would be very welcome. We believe that the cataloguing of cross-linguistically

common – and less common – pathways of change is a good first step in this direction, so we decided to supplement the overviews by Bybee et al. (1994) and Kuteva et al. (2019) with a collection of likely diachronic sources of habitual markers.

To the original 102-language sample we added a convenience sample of 85 languages for which we had information on one or more dedicated habitual markers. These were surveyed in the same way as the original variety sample. For all markers we looked for information about the likely source meanings in the surveyed grammars; in some cases we have also consulted etymological dictionaries, articles, or other sources. The two samples added together, 187 languages in total, resulted in a collection of 184 different habitual markers. The full list of these can be found in the appendices. An overview of some of the markers and their likely origins is given in Table 4.⁹ The languages in each entry are listed in alphabetical order. The likely source meanings are represented by small capitals, which in some cases cover a variety of possible English translations. For instance, the habitual marker =*taʔ* in Burmese (MYA; Sino-Tibetan, Burmo-Qiangic) is also a verb with the broad meaning KNOW, for which the sources also give translations like ‘be clever’, ‘be skilled’, ‘be capable’, ‘know how to’, or simply ‘can’.

Meaning	Language(s)
ALWAYS/STILL	Hup <i>bíg</i> (- <i>bi</i>) ‘long time’, Pennsylvania Dutch English <i>still</i> , Tapieté - <i>pi</i> ‘always’
BE ABLE	Krio <i>kin</i> , Muklomic <i>jün</i> ³ -
BE ACCUSTOMED	Slovak <i>zvyknúť</i>
BE GOOD	Ambel (Waigeo) <i>hey</i> , Ma’ya <i>fī</i> ₃ , Matbat <i>fī</i> ³
BE SAME	Pnar <i>ǵoʔ</i>
BE STRONG	Abawiri <i>kare</i> , Mandan <i>sīh-</i> , Mungbam <i>gbâbə</i> , Ot Danum (Kadorih) <i>baas</i>
BE USUAL	Muna <i>nea</i> , Papuan Malay <i>biasa</i>

⁹ For ISO codes and genealogical information on the languages in Table 4, see the Appendices. We have not included habitual markers whose diachronic origin is only tentatively suggested in the source or where a specific source lexeme cannot be identified (cf. e.g. the Kwaza construction in (11)). As already mentioned above, we have limited ourselves to likely lexical sources. Some possible grammatical sources, such as a nominalizer in Yagua (Payne & Payne 1990: 282) or a “rhetorical question marker” in Mosetén-Chimané (CAS; isolate, Bolivia; Sakel 2004: 263), are also suggested by some authors. In such cases, however, the directionality of the change is generally much less certain, and we have decided not to include them in Table 4 (though they are of course mentioned in the Appendices). An investigation of non-lexical sources of habitual markers would certainly be worthwhile, but we believe this requires work on historical data rather than a cross-linguistic grammar survey.

Meaning	Language(s)
BEFRIEND	Pakaásnovos (Wari') <i>txipa'</i>
BELONG	Cornwall English <i>belong to</i> , Krio <i>blant</i>
CARE FOR	Danish <i>pleje</i> , German <i>pflügen</i> , Western Yiddish <i>fleg</i>
CARRY	Northern Paiute <i>-(?)yakwi</i> , older Basque <i>eroan/eraman</i>
CONTINUE	Lavukaleve <i>me</i>
DO	Anamuxra <i>d-</i> , Golin <i>ere-</i> , Irish English <i>do</i> , Kulsab <i>du</i> , Kwomtari <i>a</i> , Manat <i>-rha</i> , Muyu <i>kemi</i> , Palenquero <i>asé</i> , Southern Nisu N <i>m</i> ³³
EXIST/STAY	Bora <i>i*k^ha</i> 'be', Chachapoyas Quechua <i>kawsa</i> 'be, live', Cofán (A'ingae) <i>kanse</i> 'live', Koro <i>ru/ri</i> 'stay', Lezgi <i>ɣun</i> 'be', Maldivian (Dhivehi) <i>uḷenī</i> 'live, exist', Mositén-Chimane <i>bae'i-</i> 'be, sit', Mungbam <i>nà</i> 'sit, stay', South Central Dinka <i>ɛe</i> , Yau-Nungon <i>it-</i> 'be'
GO	Cocama-Cocamilla <i>ukua</i> 'go around', Gants <i>kida</i> 'walk', Gunwinggu <i>re</i> 'go', Seze <i>jó:-</i> , Sirva <i>kida</i> 'walk', Southern Paiute <i>-mi(ya)</i> 'walk, go', Yace <i>aga</i> 'wander'
HAVE A LOT	Central Alaskan Yupik <i>-tu-</i>
KNOW	Bizkaian Basque <i>jakin</i> , Burmese = <i>ta?</i> , Cofán (A'ingae) <i>atesû</i> , Mositén-Chimane <i>chhiü-</i> , Papuan Malay <i>taw</i> , Serbo-Croatian <i>znati</i> and <i>um(j)eti</i> , Teanu <i>ovei</i> , Tok Pisin <i>save</i> , Vitu <i>lala</i>
LIKE/LOVE	Hualapai <i>sawál</i> , Khmu <i>kù?</i> , Lao <i>mak1</i> , Thai <i>chǎp</i>
PASS	Mositén-Chimane <i>jyi-</i>
THROW	Kulsab <i>kev</i>
USE	East Danish <i>brøggæ</i> , English <i>used to</i> , Swedish <i>bruka</i>
WITH	Seeku <i>wɛ̃</i>

Table 4: Likely lexical sources of habitual markers.

It is clear from Table 4 that certain patterns are attested in several unrelated languages across the world, and some of these have already been mentioned in earlier cross-linguistic studies. Bybee et al. (1994: 154–155) point to the development of habituais from verbs meaning 'know' and 'live', the latter of which corresponds to our broader grouping EXIST/STAY (i.e. verbs expressing existence or location). They also note that habitual markers sometimes develop from verbs with "lexical meaning that is very close to the grammatical meaning of habitual" (Bybee et al. 1994: 155), e.g. 'be accustomed to' or 'follow a custom'; we have also found a few examples of such markers, listed under BE ACCUSTOMED and BE USUAL.¹⁰ Kuteva et al. (2019) also mention KNOW as

¹⁰ The English *used to* habitual is also thought to have developed through such a meaning ('follow a custom') from its original meaning 'use' (see Neels 2015: 185). We have only found examples of a

a possible lexical source of habitual markers, along with GO, USE, and several existential and locational meanings corresponding to our EXIST/STAY group. At several points, the authors suggest that such existential or locational verbs often develop into habitual markers through an intermediate progressive stage (see Kuteva et al. 2019: 208, 261, 347). We can neither confirm nor refute this hypothesis on the basis of our survey.¹¹

The survey turned up a number of likely source meanings which – as far as we are aware – have not been discussed in the cross-linguistic literature. Some of these are found scattered across the world and surely reflect independent developments, whereas others are most likely historically connected. We will discuss a few examples of each category.

5.2. Transitive motion

In addition to intransitive motion verbs – ‘go’, ‘walk’, etc. – we have found a few instances of transitive motion verbs as likely sources of habitual markers, namely verbs meaning THROW and CARRY. One is the Papuan language Kulsab (Kursav). Daniels (2015: 923–924) mentions three habitual strategies in this language: a dedicated habitual suffix *-di*, which appears to be infrequent on its own, and two serial verb constructions where the habitual suffix is added to the serial verb; one of these constructions uses the verb *du-* ‘do’, the other is formed with *kev-* ‘throw’, as in (13):

(13) Kulsab (FAJ; Nuclear Trans New Guinea, Madang; Daniels 2015: 924)

Mot gapira, sarua ka vu kevi-d-o.
day all work MD get throw-HAB-3PL
‘Every day, they do the work.’

habitual from a verb meaning ‘use’ in Northern Europe. Kuteva et al. (2019: 450) mention only English *used to* as an example of this pathway.

¹¹ Note that Kuteva et al. (2019: 249) also seem to suggest that the development from KNOW to habitual (through BE ABLE) is especially frequent “[i]n pidgin and creole languages”. Based on our survey we would rather say that the development is cross-linguistically widespread in general, not just in pidgins and creoles.

Another example is the Northern Paiute habitual marker *-(?)yakwi* as illustrated in (14a), which almost certainly derives from the transitive motion verb *yakwi* ‘carry’ (Thornes 2003: 410–411, 416–418), as shown in (14b):

(14) Northern Paiute (PAO; Uto-Aztecan, Numic; Thornes 2003: 186, 542)

- a. *yaisi* *owi-u* *yotsi-u-ga-ʔyakwi* *tui-haʔu-tui*
 then DEM-SUFF fly.SG-PUNCT-TRNSL-HAB any-how-any
 ‘... and it [a mythical creature] would fly off from there in every direction’
- b. *uuni-ku* *ti-ma-yakwi-na* *su = ti-woitsami-di* *ka = tokaano*
 that.kind-OBL AP-IP(hand)-carry-PTCP NOM = AP-watchman-NMLZ OBL = night
 ‘That’s what he was carrying, that night watchman’

Similarly, in their grammar of Basque (EUS; isolate, Spain/France), Hualde & Ortiz de Urbina (2003: 296) mention that “[o]lder forms of the language” had a habitual construction with the verb *eroan/eraman* ‘carry’. According to the authors, this habitual use of ‘carry’ was restricted to transitive predicates, whereas intransitive predicates formed the habitual construction with *joan* ‘go’.¹² We have not found any other examples of habitual markers from transitive motion verbs, but given that the three markers mentioned here are certainly not the result of language contact, it may be a more common pathway than our three examples suggest.

5.3. BE STRONG

In four languages in the convenience sample, lexemes meaning BE STRONG are reported to have developed habitual meaning. In Mandan (MHQ; Siouan), Mixco (1997: 27, 52) reports the use of the verb *sîh-* ‘be strong’ as a habitual auxiliary. For Mungbam (MIJ; Atlantic-Congo, Volta-Congo), Lovegren (2013: 235) mentions a habitual serial verb construction with the verb *gbâbâ* ‘be strong’, and in Ot Danum (Kadorih; OTD; Austro-nesian, Malayo-Polynesian), Inagaki (2013) describes a habitual marker derived from the adjective *baas* ‘strong’. Finally, Yoder (2020: 320) suggests that the postverbal habitual marker *kare* in Abawiri (FLH; Lakes Plain, East Lakes Plain) “likely developed

¹² While Basque was one of the languages in our variety sample, we did not include the constructions with *joan* ‘go’ and *eroan/eraman* ‘carry’ in our initial survey (Section 4), as these are reported to be obsolete in the modern language.

from the adjective *kare* ‘strong, hard’, where it was used attributively after a nominalized verb”. Although the markers in Abawiri and Ot Danum might reflect a more general areal tendency, the Mandan and Mungbam constructions are certainly the result of independent developments. It would be interesting to investigate whether other instances of this pathway can be found and whether it developed through an intermediate stage BE ABLE or in some other way.

5.4. BE GOOD

Another pattern which we have only found in a few languages is the use of a predicate meaning BE GOOD as a habitual construction. Unlike transitive motion and BE STRONG, this has so far only been attested in one specific area, namely the Halmahera and Raja Ampat archipelagos in eastern Indonesia. Arnold (2018: 550) discusses the use of *hey* ‘be good’ as a habitual marker in Ambel (Waigeo; WGO) and points to parallels in Matbat (XMT; Remijsen 2010) and Ma’ya (SLZ; van der Leeden n.d.). These three languages all belong to the same subbranch of Austronesian (Raja Ampat-South Halmahera), so the development could conceivably be a case of shared inheritance. However, Arnold (p. c.) notes that a similar use of a lexeme meaning ‘good’ appears to be attested in the unrelated language Tidore, though this is not discussed as a habitual construction by van Staden (2000); consider the example and van Staden’s translation in (15).

(15) Tidore (TVO; North Halmahera, Northern North Halmahera; van Staden 2000: 319)

<i>mina</i>	<i>gahi</i>	<i>ngam</i>	<i>laha</i>
3SG.F	make	food	good
‘She can cook and does it often’			

If this use of *laha* in Tidore indeed qualifies as a habitual construction, it would suggest that at least some languages in the area may have developed habituals from a lexeme meaning ‘(be) good’ through language contact. To confirm this, however, further documentary and descriptive work on the languages of Halmahera and Raja Ampat is necessary.

5.5. BELONG

Finally, we will consider habitual markers from a verb meaning BELONG. This is attested in two languages in our convenience sample, the English dialect of Cornwall in southwestern England and Krio, an English-lexifier creole. The Cornwall English construction is attested both with infinitive and gerund complements, as shown in (16).

(16) Cornwall English (ENG; Indo-European, Germanic; EDD, s.v. *belong*, sense 4)

- a. *I am not so ill as I belong to be.*
'I am not so ill as I generally am.'
- b. *I don't belong going to Church, but I will this once.*
'I don't usually go to church, but I will this once.'

The Krio habitual, called a “customary action marker” by Hancock (1971: 127), has the form *blant*, which as an independent verb means ‘belong’ and historically derives from English *belong to*. Yillah & Corcoran (2007: 180–181) report that some speakers only use habitual *blant* in past contexts, contrasting it with present habitual *kin*, whereas for other speakers *blant* is an “emphatic” habitual; compare (17a) and (17b).

(17) Krio (KRI; Indo-European, Germanic; Yillah & Corcoran: 181)

- a. *Kabba blant kam*
K. HAB.PST come
'Kabba used to come'
- b. *Olu blant go London fo Krismes*
O. HAB.EMPH go London for Christmas
'Olu always goes to London for Christmas'

In this case, as in the case of the BE GOOD habituals in eastern Indonesia, there is very likely a historical connection between the constructions. According to Hancock (1971: 122–126), southwestern dialects of English were among the main sources of the Krio lexicon, and habitual *blant* is identified as one such “dialect-derived” item (see Hancock 1971: 127–128). Hancock notes that it is also attested in Gullah (GUL), another

English-lexifier creole with significant southwestern input (see Hancock 1994; McWhorter 2019), although Rickford (1974: 101–102) characterizes the construction as obsolete in modern Gullah. At any rate, it seems likely that the use of *BELONG* as a habitual marker was a dialectal English development which was later transferred to Krio and Gullah. A closer study of the construction in traditional Cornwall English would thus also be of great interest to the history of English-lexifier creoles.

6. Discussion

6.1. Prevalence and form of habituals

The findings from our 102-language variety sample suggest that dedicated habitual markers are cross-linguistically common, being attested in 52 of the languages in the sample. Affixes were the most common type of marker in this sample. These findings may at first glance seem surprising in light of some earlier work, so it is worth briefly reflecting on what they may or may not tell us.

First, consider Dahl's (1985: 96) finding that habituality is only a "major TMA category" in a few languages in his survey. This might initially seem at odds with our finding that dedicated habitual markers are cross-linguistically common. However, it is important to keep in mind that Dahl's questionnaire-based study used a very specific "and indeed somewhat arbitrarily" defined notion of "major TMA category" (Dahl 1985: 52). A tense-mood-aspect form in an individual language was only classified as "major" if it occurred in at least six of the translations in Dahl's questionnaire and was attested in non-negated declarative main clauses. As Dahl (1985: 96) explicitly notes in connection with habitual markers, it is "fairly plausible that similar categories can be found in other languages" even if they are too infrequent to qualify as "major" in his terms. By contrast, we have not applied any frequency criteria in our survey, but simply asked whether one or more habitual markers are attested in the languages in the sample (cf. also the point that "target uses are not all of the same kind" in Kuteva et al. 2019: 8). Our method is thus not directly comparable to Dahl's and should not be taken to contradict his conclusions. Consider the Japanese past habitual construction with the nominalizer *mono* in (18) (see also (10) above):

(18) Japanese (own data; F22 from Kantō region)

hebi = o mikakeru to, ishi = o hiro-tte hebi = ni
 snake = ACC catch.sight when stone = ACC pick.up-CONJ snake = DAT

muka-tte nage-ta mono da-tta yo
 go.towards-CONJ throw-PST NMLZ COP-PST AFF

‘When I saw a snake, I would pick up a stone and throw it at the snake.’

This habitual construction is clearly part of the Japanese grammatical system in the sense that speakers understand it and can produce examples of it when prompted, but it appears to be marginal in the contemporary vernacular language.¹³ It is thus no surprise that it did not make it into Dahl’s overview of major TMA categories in Japanese (Dahl 1985: 158), and it is quite possible that some of the other markers in our survey would not qualify as major tense-mood-aspect markers if investigated with Dahl’s questionnaire. One way to control for such frequency differences in future work would be to study habitual markers in a cross-linguistic corpus, thus allowing for a comparison of usage frequencies between languages.

On a similar note, Thieroff (2000: 295) concludes that “there are only a few languages with a commonly accepted habitual category” in his sample of European languages. Again, however, the findings from our variety sample cannot be directly compared to this earlier survey. First, because we used a global sample controlling for genealogical affiliation, we only surveyed a small number of European languages and cannot make any conclusions about the languages of Europe in general.¹⁴ Second, Thieroff only includes “fully grammaticalized” markers in his survey, described as forms “obligatory in the appropriate contexts, and hence [...] belonging to the verbal

¹³ The consultant who produced (18) (as a translation of sentence B7 in Dahl’s TMA questionnaire) immediately remarked that it sounds literary and is probably not used much in the spoken language; our two other native-speaker consultants did not use the construction at all. Rie Obe (p. c.) confirms to us that the construction sounds old-fashioned and belongs to a formal or literary register.

¹⁴ For what it is worth, we found a dedicated habitual marker in two of the five European languages in our variety sample – Basque and Georgian (KAT; Kartvelian) – while none was found in the other three: Abkhaz (ABK; Abkhaz-Adyge), Sign Language of the Netherlands (DSE; LSFic), and Turkish (TUR; Turkic). Note that Thieroff’s sample is heavily skewed: 33 of the 40 investigated languages are Indo-European, of which 11 are Germanic and 10 Slavic. As the author himself notes, we think correctly, the results “probably would look less homogeneous if more languages, especially more non-Indo-European languages, were investigated” (Thieroff 2000: 267).

paradigm” (Thieroff 2000: 296). Swedish (SWE) *bruka* and German (DEU) *pflügen* are mentioned as habitual expressions which do not satisfy this criterion. By contrast, we have not applied any criterion of “full” grammaticalization in our study, as it is very difficult to operationalize in a survey of descriptive grammars. Note, however, that most habitual markers found in our variety sample belong to formal categories usually regarded as grammatical, e.g. affixes, clitics, and particles. Some of the auxiliaries in the survey are also explicitly described as grammatical in the sources,¹⁵ and we think it is fair to conclude that most of the 75 habitual markers in the variety sample can be considered grammatical. Even if grammaticalized habitual markers may be rare in European languages, our survey thus suggests that they are quite widespread outside of Europe. Especially noteworthy is their frequent attestation in South America (22 out of 33 languages in the variety sample), a finding which deserves further attention in future studies.

Finally, we will briefly comment on the relatively high frequency of affixes that we found in the variety sample (see Section 4.2), in particular because Bybee et al. (1994: 20) suggest that habitual meanings “typically have periphrastic expression”. At first glance, our survey seems to contradict this and indicate that dedicated habituals tend to be bound morphemes. However, a high frequency of bound morphemes may also to some extent be an artefact of the method of surveying descriptive grammars and grammar sketches: while such works can be expected to describe most or all inflectional morphemes in a given language, periphrastic constructions are probably more liable to be overlooked, in particular constructions with lower frequency. All else being equal, periphrastic constructions may thus also more easily be overlooked in a grammar survey, and for this reason, we do not think one can conclude that habituals in general tend to be expressed by bound morphemes. On the other hand, however, there is no indication in our survey that periphrastic habituals are especially frequent (*pace* Bybee et al. 1994).

6.2. Habituals and pluractionals

As noted above (see Section 2), in some languages pluractional markers may be used to express habitual meaning, and many linguists describe habituality as a subtype of pluractionality. Pluractional constructions in the languages of the world have recently

¹⁵ See e.g. Vallejos (2016: 372–379) on Cocama-Cocamilla (COD; Tupian, Tupi-Guarani) and Hengeveld & Fischer (2018) on Cofán/A’ingae.

been surveyed by Mattioli (2019, 2020), and it is worth pointing out a few similarities and differences between his and our findings.

Mattioli (2019: 20–42) distinguishes between ‘core’ and ‘non-prototypical’ functions of pluractional markers. The core functions are iterativity (repetition of an event on a single occasion), frequentativity (repetition on multiple occasions), spatial distributivity (repetition in different places), and participant plurality. Habituality, which for Mattioli involves the combination of repetition and typicality, is one of the many non-prototypical (or ‘additional’; Mattioli 2020) functions of pluractional markers. Note that a marker is only counted as a pluractional if it has at least one of the core functions; markers limited to one or more non-prototypical functions, such as habituality, are not included in the survey (see Mattioli 2019: 19–21). The habitual expressions included in Mattioli’s survey can thus be regarded as pluractional markers which ‘extend’ into the semantic domain of habituality.

Three expression formats are found to be prevalent for pluractional markers by Mattioli (2019: 65–82): affixes, reduplication, and lexical alternation. Because pluractional markers may be used in habitual contexts in some languages, these strategies may thus also be found with a habitual interpretation. For instance, verb reduplication in Paiwan may receive a habitual interpretation, as in (19a), but also has core pluractional functions like iterativity, as in (19b) (Chang 2006: 53–56; Mattioli 2019: 60, 206).

(19) Paiwan (PWN; Austronesian; Chang 2006: 54–55)

- a. *siu~siup ti zepul nu siaw*
 suck<AV> ~PLAC NOM.SG Zepul IRR.TEMP soup<AV>
 ‘Zepul sucks it when she eats soup.’
- b. *v<en>aridi~ridi ti kulele ta zua ?utubay nimadu*
 turn<AV> ~PLAC NOM.SG Kulele OBL that motorbike 3SG.GEN
 ‘Kulele kept turning his motorbike.’

Similarly, the ‘plural-action’ stems of verbs in Maricopa (MRC; Cochimi-Yuman) are said to indicate “that the activity expressed by the verb is repeated or habitual or that the object of the verb is plural” (Gordon 1985: 99–100). For dedicated habitual markers, by contrast, we found no examples of lexical alternation, and reduplication is very rare in our material, being found only in a single language in the variety sample

and three languages in the convenience sample, always in combination with some other marking.¹⁶ These expression formats thus appear to be marginal or non-existent for dedicated habituais, and our prediction would be that if habituality in a given language can be expressed through lexical alternation or reduplication alone, this construction will also have ‘core’ pluractional functions.

Mattiola (2019: 149–157) points to four possible lexical sources of pluractionals in his data. These are demonstratives (in Chadic languages), verbs of liking, verbs of position or location, and motion verbs. The last three are also attested multiple times in our material (see the categories LIKE/LOVE, EXIST/STAY, and GO in Section 5). The fact that we have gathered a longer list of potential sources of dedicated habituais should be no surprise, as our convenience sample was explicitly designed to identify as many of these as possible. Moreover, in the case of pluractionals expressed through reduplication or lexical alternation, it is of course not possible to point to any individual lexical source.

6.3. *Past habituais*

It has been observed that there seems to be a special relation between habitual meaning and past tense, in the sense that if a habitual marker is limited to particular tense contexts, these are more likely to be past than non-past tenses (see e.g. Dahl 1985; Bybee et al. 1994: 151–154; Carlson 2012: 838). Two different types of past habituais may be distinguished: unanalyzable portmanteau markers which simultaneously express past and habitual meaning, and habitual markers which obligatorily cooccur with past tense markers. Compare the Betta Kurumba portmanteau past habitual marker *-ott* described by Coelho (2003) with the Northeast Maidu habitual marker *-us*, which according to Shipley (1963) obligatorily occurs with the past tense suffix *-ʔ*:

(20) Betta Kurumba (XUB; Dravidian, South Dravidian; Coelho 2003: 45)

pin aḍəŋ atni:ri alka pəyottla

pina aḍən-ka atni:ri alka pəy-ott-il-a

then 3SG.REM-DAT onwards there.DAT go-HAB.PST-NEG-3SG

‘Then, from that day on, she never went there.’

¹⁶ The language in the variety sample is Kwaza, cf. (11), while the three languages from the convenience sample are Golin (GVT), Korowai (KHE), and Yagaria (YGR) (all Nuclear Trans New Guinea). In these three languages, habituality is reported to be expressed by a combination of reduplication and an auxiliary.

(21) Northeast Maidu (NMU; Maiduan; Shipley 1963: 52)

péne-m ník-kotò-c'o-m máj-c'o-m ʔyhéj-ʔ-us-a-n.
 two-NOM 1POSS-grandmother-DU-NOM 3NSG-DU-NOM accompany-PST-HAB-IND-3
 'My two grandmothers used to go with us.'

In the sample used by Bybee et al. (1994), only the second type was found, i.e. past habitual markers that always cooccur with overt past tense marking. Bybee et al. suggest a functional motivation for the restriction to past contexts in some languages: whereas habitual meaning is one of the default readings of simple present tenses, simple past tense forms are more likely to be interpreted as perfective. In other words, according to the authors, there is a stronger pressure towards explicit habitual marking in the past than in the present (Bybee et al. 1994: 153–154).

We also encountered past habituals while surveying our 102-language sample and decided to investigate the frequency of such markers. For each marker in the sample, we noted whether it was described as restricted to past-time reference and whether it was in paradigmatic opposition to one or more non-past habitual markers. The languages with habitual markers could then be grouped into three types: those where only a past habitual was recorded, those with an opposition between a past and a non-past habitual, and those without tense restrictions or contrasts in their habitual marker(s), i.e. where the same marker may occur in either past or non-past contexts. The figures are given in Table 5.

Habitual–past relation	No. of languages
Habitual limited to past	3
Past ≠ non-past habitual	4
Habitual without tense restriction	45
No habituals found	50
Total	102

Table 5: Past and non-past habituals in the variety sample.

Perhaps surprisingly, dedicated past habituals turned out to be fairly infrequent in the sample, being attested in only seven out of the 52 languages where we found at least one habitual marker.¹⁷ While languages restricting a habitual marker to past contexts

¹⁷ Of these, portmanteau past habitual markers were found in Berbice Dutch Creole (BRC; Indo-European, Germanic), Betta Kurumba, Japanese, Marind (MRZ; Anim), and Wambaya (WMB; Mirndi). Ket

are thus clearly found, this does not seem to be particularly frequent in the languages surveyed here. However, note that we have only counted a marker as a past habitual if the sources explicitly described it as restricted to past contexts. We thus cannot rule out that some such markers have been overlooked. Furthermore, there may also be languages where a habitual marker is more frequently – but not exclusively – used in past contexts (for two examples, see Olawsky 2006: 465–467; Vallejos 2016: 373–375). Again, we think that a promising way forward would be a corpus-based typological study allowing for the comparison of usage patterns across languages.

6.4. Subtypes of habitual meaning

In most of the cross-linguistic literature on aspect, habitual meaning is treated as a single semantic category (e.g. Comrie 1976; Dahl 1985; Bybee et al. 1994). On the basis of his survey, Carlson (2012: 842) concludes that languages with a grammatical habitual “seem to introduce just one marker of “habituality,” and nothing resembling a field of contrasting markers”. While this indeed seems to be the case also in most of the languages we have surveyed, we think the picture may be slightly more complicated. In a few of the languages where we found more than one habitual marker, the sources either explicitly state or hint at semantic contrasts between them. Hence, at least for some languages it seems to be necessary to distinguish between several different subtypes of habitual meaning.

Consider the case of Amur Nivkh as described by Nedjalkov & Otaina (2013). Here a distinction is made between what is termed a *USITATIVE* and a *QUALITATIVE HABITUAL* in the grammar. Whereas the *usitative*, illustrated in (22a), “denotes a repeated regularly performed action” and may occur with non-human subjects, verbs with the *qualitative habitual* suffix “characterize the subject (human) as inclined to perform an action [...] habitually” (Nedjalkov & Otaina 2013: 94, 139). The *qualitative habitual*, illustrated in (22b), is also said to be of more limited productivity than the *usitative*.

(22) Amur Nivkh (NIV; Nivkh; Nedjalkov & Otaina 2013: 95, 139)

- a. *Tolf* *if* *vi-řa-đ*.
 summer he go-USIT:3SG-IND
 ‘In the summer he usually goes [to that village]’

(KET; Yeniseian) is an intermediate case, as past tense is both marked on the verb and with the portmanteau habitual past particle *ba* (Vajda 2004).

- b. *Ōla va-xə-ḡ.*
 child fight-QUAL.HAB-IND
 ‘The child often fights’

Comparable distinctions have recently been described for Plains Cree (Wolvengrey 2026) and Slovak (Genis & Kyselica 2026). In many cases of multiple habitual markers, however, the semantic difference between them (if any) is less clear. The two habitual auxiliaries in Cofán/A’ingae, derived from verbs meaning KNOW and LIVE, are syntactically distinct, as the former requires an infinitival complement, the latter a same-subject form; see the examples in (23). Whether there is any semantic difference between them, however, is unclear (Kees Hengeveld, p. c.).

(23) Cofán/A’ingae (Fischer & Hengeveld 2023: 104)

- a. *munda = ma an-ye atesû = ti = ki?*
 peccary = ACC.REAL eat-INF HAB = Q = 2
 ‘Do you eat peccary?’
- b. *ja’ñu kundase = pa kanse = mbi = ’ya*
 now tell = SS HAB = NEG = ASS
 ‘Nowadays I don’t tell stories.’

Another such case is Pawnee (PAW; Caddoan) as described by Parks (1976). Here a distinction is made between a “habitual” suffix *-uku* and a “usitative” preverb *ut-*. The habitual is said to describe “[a]ctivities which occur according to habit or occur customarily”, while the usitative is used for “activities which are continuous, habitual, or usual; also activities which occur now and then” (Parks 1976: 199–200). From the descriptions alone, the usitative might appear to have a more general imperfective meaning, but based on the examples in the grammar, it seems that both constructions are to be characterized as dedicated habituals in our terms. The exact difference between them, however, remains unclear to us. We think such languages with multiple habitual markers – and the possible semantic and pragmatic differences between them – deserve more attention in future work.

6.5. Negative habituals

Finally, we wish to briefly mention a phenomenon which we have not seen any references to in the cross-linguistic literature. In some languages, a distinct (portman-teau) negative habitual marker is attested, expressing that a situation does not usually occur. This was found in four languages in the variety sample, including Amur Nivkh and Cofán/A'ingae:

(24) Amur Nivkh (Mattissen 2003: 200)

əyrəku ñivɣ-gu tʃaj-χavu-t ra-ysu-ta.
 former_times person-PL tea-heat-CV:3PL drink-NEG.HAB-ENU:3PL
 ‘In the past, the Nivkh did not heat tea and drink it.’

(25) Cofán/A'ingae (Fischer & Hengeveld 2023: 98)

Je'nda kûti'chu kuku = ta = ti tsa = 'ka = en fi'thi = masia?
 then tree devil = NEW = Q ANA = SIM = ADVZ kill = NEG.HAB
 ‘Then the *yachapo* demon is not killed like that?’

In addition, a dedicated negative habitual is found in Coptic, one of the languages in the convenience sample:

(26) Sahidic Coptic (COP; Afroasiatic, Egyptian; van der Vliet & Zakrzewska 2026: 64)

p-refčioue me = f-ei etbe laau eimētei čekas
 DEF.SG.M-thief NEG.HAB = 3SG.M-come for anything unless in.order.that

e = f-ehōbt

FUT.3 = 3SG.M-steal

‘The thief does not come for any other reason than to steal’

In all of these five languages, the negative habitual exists alongside a positive habitual marker. We have thus found no examples of languages with a grammatical negative habitual without a positive habitual, but we of course cannot rule out that such a language may exist. We also have limited information on the historical development of negative habituals, the only relatively certain source being a verb meaning ‘ignore,

not know’ in Amur Nivkh (on which see Mattissen 2003: 199–200). More research into the origins of this type of marker would thus be welcome.

7. Conclusions

This paper has presented a cross-linguistic study of habitual markers, surveying their formal diversity and possible lexical sources around the world. To sum up, we present our main conclusions point for point, broadly corresponding to the questions we asked at the outset of the investigation (see Section 3):

1. In 52 of the 102 languages in our variety sample, at least one dedicated habitual marker was found. In 12 of these languages, more than one marker was found. Dedicated habitual markers are thus cross-linguistically common, and it is not exceptional for a language to have more than one such marker.
2. The most common type of marker in the variety sample was affixes, followed by auxiliaries and particles. Less frequent formal types in the surveyed languages include clitics, serial verbs, and other complex constructions.
3. The survey of lexical sources suggests that a very large number of different linguistic items may develop into habitual markers. We pointed to some more or less cross-linguistically common pathways, noting that some of the rarer ones may be indicative of language contact.
4. We found dedicated past habituais to be relatively infrequent, being attested in only seven languages in the variety sample. We also pointed to other possible distinctions in the domain of habituality, noting that some languages distinguish between more than one subtype of habitual meaning or between positive and negative habitual markers.

We hope that these cross-linguistic generalizations will be of interest to typologists as well as historical and descriptive linguists, and we are convinced that many interesting habitual constructions still remain to be described across the world’s languages.

Acknowledgements

We are grateful to everyone who replied to our LingTyp query (17 Apr 2023) for their many helpful suggestions and comments. Thanks are also due to audiences in Helsinki, Paris, and Singapore, the two anonymous reviewers, and the members of the research groups “Habituals” and “Serial Predicates Across Modalities” (University of Amsterdam). Finally, the first author wishes to thank Rie Obe, Shion Sato, Shinnosuke Murakami, and the other students at the University of Osaka who shared their knowledge of Japanese with him during his research stay in May and June 2024.

Abbreviations

1 = 1st person	DET = determiner	OBL = oblique
2 = 2nd person	DU = dual	PL = plural
3 = 3rd person	DYN = dynamic	PLAC = pluractional
ACT = active	EMPH = emphatic	PLS = plural subject
ADD = additive focus	ENU = enumerative	POSS = possessive
ADJZ = adjectivizer	ERG = ergative	PST = past
ADVZ = adverbializer	FREQ = frequentative	PTCP = participle
AFF = affirmative	FUT = future	PUNCT = punctual
ANA = anaphoric	HAB = habitual	Q = question (interrogative)
AP = antipassive	HUMPL = human plural	QUAL = qualitative
APPL = applicative	IND = indicative	REAL = realis
ASS = assertive	INV = inverse	REM = remote
AV = actor voice	IP = instrumental prefix	RPT = reportative
CNJ = conjunct	IPFV = imperfective	S = intransitive subject
CNTR = contrastive topic	IRR = irrealis	SG = singular
COMP = complementizer	LOC = locative	SIM = similitive
CONJ = conjunctive	M = masculine	SS = same subject
COP = copula	MD = middle deictic	SUFF = suffix
CV = converb	NEG = negative	TEMP = temporal
DAT = dative	NEW = new topic	TOP = topic
DECL = declarative	NMLZ = nominalizer	TRNSL = translocative
DEF = definite	NOM = nominative	USIT = usitative
DEP = dependent	NSG = non-singular	YM = Yuchi male

Appendices (available at <https://doi.org/10.60923/issn.2785-0943/20272>)

1. Habitual markers in the 102-language variety sample.
2. Habitual markers in the 85-language convenience sample.

References

The bibliography contains all sources used for the survey as well as those cited in the article. All hyperlinks were last checked on 18 April 2025.

- Abbi, Anvita. 2013. *A grammar of the Great Andamanese language*. Leiden: Brill.
- Andersen, Torben. 2007. Auxiliary verbs in Dinka. *Studies in Language* 31(1). 89–116. <https://doi.org/10.1075/sl.31.1.04and>
- Anderson, Vicki Michael. 2014. *Bidialectalism: An unexpected development in the obsolescence of Pennsylvania Dutchified English*. Durham, NC: Duke University Press.
- Andrews, Kenneth Ralph. 1994. *Shawnee grammar*. Columbia: University of South Carolina. (Doctoral dissertation.)
- Arnold, Laura. 2018. *A grammar of Ambel: An Austronesian language of Raja Ampat, west New Guinea*. Edinburgh: University of Edinburgh. (Doctoral dissertation.) <http://hdl.handle.net/1842/31120>
- Årsjö, Britten. 1999. *Words in Ama*. Uppsala: Uppsala University. (MA thesis.)
- Audring, Jenny, Sterre Leufkens & Eva van Lier. 2021. Small events: Verbal diminutives in the languages of the world. *Linguistic Typology at the Crossroads* 1(1). 223–256. <https://doi.org/10.6092/issn.2785-0943/13427>
- Bakker, Dik. 2011. Language sampling. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 100–127. Oxford: Oxford University Press.
- Bar-el, Leora. 2008. Verbal number and aspect in Skwxwú7mesh. *Recherches linguistiques de Vincennes* 37. 31–54. <https://doi.org/10.4000/rlv.1695>
- Bellamy, Kate. 2018. *On the external relations of Purépecha: An investigation into classification, contact and patterns of word formation*. Leiden: Leiden University. (Doctoral dissertation.)
- Bendor-Samuel, John Theodor. 1981. *The structure and function of the verbal piece in the Jebero language*. Lima: Instituto Lingüístico de Verano.
- Bertet, Denis. 2020. *Aspects of Tikuna grammar (San Martín de Amacayacu variety, Colombia): Phonology, nominal phrase, predicative phrase*. Lyon: Université de Lyon. (Doctoral dissertation.)
- Berry, Keith. 1995. *A description of the Abun language: Phonology and basic morphosyntax*. Melbourne: LaTrobe University. (MA thesis.)
- Bertinetto, Pier Marco. 2014. Ayoreo. In Mily Crevels & Pieter Muysken (eds.), *Lenguas de Bolivia*, vol. 3, 369–413. La Paz: Plural Editores.

- Bertinetto, Pier Marco & Alessandro Lenci. 2012. Habitual and generic aspect. In Robert I. Binnick (ed.), *The Oxford handbook of tense and aspect*, 852–880. Oxford: Oxford University Press.
- Blackings, Mairi & Nigel Fabb. 2003. *A grammar of Ma'di*. Berlin: Mouton de Gruyter.
- Bliss, A. J. 1972. Languages in contact: Some problems of Hiberno-English. *Proceedings of the Royal Irish Academy: Archaeology, Culture, History, Literature* 72. 63–82. <https://www.jstor.org/stable/25506261>
- Boneh, Nora & Łukasz Jędrzejowski. 2019a. *Habituality from a typological and diachronic perspective*. Special issue of *STUF – Language Typology and Universals* 72(1). <https://www.degruyterbrill.com/journal/key/stuf/72/1/html>
- Boneh, Nora & Łukasz Jędrzejowski. 2019b. Reflections on habituality across other grammatical categories. *STUF – Language Typology and Universals* 72(1). 1–20. <https://doi.org/10.1515/stuf-2019-0001>
- Bornholmsk Ordbog*, online edn. Copenhagen: University of Copenhagen. <https://bornholmskordbog.ku.dk/>
- Bowern, Claire. 2011. *Sivisa Titan: Sketch grammar, texts, vocabulary based on material collected by P. Josef Meier and Po Minis*. Honolulu: University of Hawai'i Press.
- Bradshaw, Robert L. 2012. *Doromu-Koki grammar sketch*. Ukarumpa: SIL-PNG.
- Briley, David. 1997. Four grammatical marking systems in Bauzi. In Karl J. Franklin (ed.), *Papers in Papuan linguistics no. 2*, 1–131. Canberra: ANU.
- Broadwell, George Aaron. 2006. *A Choctaw reference grammar*. Lincoln: University of Nebraska Press.
- Bromley, C. Myron. 1981. *A grammar of Lower Grand Valley Dani*. Canberra: ANU.
- Brown, Jason, Alex Muir, Kimberley Craig & Karika Anea. 2016. *A short grammar of Urama*. Canberra: Asia-Pacific Linguistics.
- Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: The University of Chicago Press.
- Caballero, Gabriela. 2022. *A grammar of Choguita Rarámuri*. Berlin: Language Science Press. <https://doi.org/10.5281/zenodo.7189161>
- Carlson, Greg. 2012. Habitual and generic aspect. In Robert I. Binnick (ed.), *The Oxford handbook of tense and aspect*, 828–851. Oxford: Oxford University Press.
- Carlson, Robert. 1994. *A grammar of Supyire*. Berlin: Berlin: Mouton de Gruyter.
- Chafe, Wallace. 2015. *A grammar of the Seneca language*. Berkeley: University of California Press.

- Chirikba, Viacheslav A. 2003. *Abkhaz*. Munich: LINCOM.
- Cleary-Kemp, Jessica. 2015. *Serial Verb Constructions revisited: A case study from Koro*. Berkeley: UC Berkeley. (Doctoral dissertation.)
- Coelho, Gail Maria. 2003. *A grammar of Betta Kurumba*. Austin: University of Texas at Austin. (Doctoral dissertation.)
- Comrie, Bernard. 1976. *Aspect*. Cambridge: Cambridge University Press.
- Coupe, Alexander R. 2007. *A grammar of Mongsen Ao*. Berlin: Mouton de Gruyter.
- Cristofaro, Sonia. 2004. Past habituals and irrealis. In Yury A. Lander, Vladimir A. Plungian & Anna Yu. Urmanchieva (eds.), *Irrealis i irreal'nost'* (Issledovaniya po teorii grammatiki 3), 256–272. Moscow: Gnosis.
- Crowell, Thomas Harris. 1979. *A grammar of Bororo*. Ithaca: Cornell University. (Doctoral dissertation.)
- Crowley, Terry. 2002. *Serial verbs in Oceanic: A descriptive typology*. Oxford: Oxford University Press.
- Curnow, Timothy. 1997. *A grammar of Awa Pit (Cuaiquier): An indigenous language of south-western Colombia*. Canberra: ANU. (Doctoral dissertation.)
- Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Blackwell.
- Daniels, Don. 2015. *A reconstruction of Proto-Sogeram: Phonology, lexicon, and morphology*. Santa Barbara: UC Santa Barbara. (Doctoral dissertation.)
- Daniels, Don. 2020. The history of tense and aspect in the Sogeram family. *Journal of Historical Linguistics* 10(2). 167–208. <https://doi.org/10.1075/jhl.18012.dan>
- de Angulo, Jaime & L. S. Freeland. 1930. The Achumawi language. *International Journal of American Linguistics* 6(2). 77–120. <http://www.jstor.org/stable/1263305>
- Desta, Girma Mengistu. 2015. *A grammar of Sezo*. Addis Ababa: Addis Ababa University. (Doctoral dissertation.)
- Desta, Girma Mengistu. 2021. Serial verb constructions in Sezo. In Derib Ado, Almaz Wasse Gelagay & Janne Bondi Johannessen (eds.), *Grammatical and sociolinguistic aspects of Ethiopian languages*, 197–212. Amsterdam: John Benjamins.
- Ding, Hongdi & Lama Qiufuyuan Ziwo. 2021. Aspect marking in Niesu, a dialect of Nuosu in Sichuan, China. *Himalayan Linguistics* 20(1). 1–27.
- Davidson, Matthew. 2002. *Studies in Southern Wakashan (Nootkan) grammar*. Buffalo: SUNY Buffalo. (Doctoral dissertation.)
- Donohue, Mark. 1999. *A grammar of Tukang Besi*. Berlin: Mouton de Gruyter.
- dos Anjos, Zoraide. 2011. *Fonologia e Gramática Katukina-Kanamari*. Amsterdam: VU Amsterdam. (Doctoral dissertation.)

- Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures online*. <https://wals.info/>
- Dunn, Michael J. 1999. *A grammar of Chukchi*. Canberra: ANU. (Doctoral dissertation.)
- EDD = Joseph Wright (ed.). 1898–1905. *English Dialect Dictionary*. 6 vols. Oxford: Oxford University Press. <http://eddonline-proj.uibk.ac.at/edd/>
- Elliott, John A. 2021. *A grammar of Enxet Sur*. Mānoa: University of Hawai‘i at Mānoa. (Doctoral dissertation.)
- Enfield, N. J. 2007. *A grammar of Lao*. Berlin: Mouton de Gruyter.
- Enfield, N. J. 2008. Verbs and multi-verb constructions in Lao. In Anthony V. N. Diller, Jerold A. Edmondson and Yongxian Luo (eds.), *The Tai-Kadai languages*, 83–183. London: Routledge.
- Ennever, Thomas. 2021. *A grammar of Ngardi*. Berlin: De Gruyter Mouton.
- Epps, Patience. 2008. *A grammar of Hup*. Berlin: Mouton de Gruyter.
- Evans, Nicholas. 1995. *A grammar of Kayardild*. Berlin: Mouton de Gruyter
- Evans, Nicholas. 2003. *Bininj Gun-Wok: A pan-dialectal grammar of Mayali, Kunwinkju and Kune*. Canberra: Pacific Linguistics.
- Evans, Nicholas, Jutta Besold, Hywel Stoakes & Alan Lee (eds.). 2005. *Materials on Golin: Grammar, texts and dictionary*. Melbourne: University of Melbourne.
- Everett, Daniel L. & Barbara Kern. 1997. *Wari’: The Pacaas Novos language of Western Brazil*. London: Routledge.
- Fedden, Sebastian Olcher. 2011. *A grammar of Mian*. Berlin: Mouton de Gruyter.
- Fehn, Anne-Maria. 2014. *A grammar of Ts’ixa (Kalahari Khoe)*. Cologne: University of Cologne. (Doctoral dissertation.)
- Fernández Garay, Ana. 1998. *El tehuelche: una lengua en vías de extinción*. Valdivia: Universidad Austral de Chile.
- Fischer, Rafael & Kees Hengeveld. 2023. A’ingae (Cofán). In Patience Epps & Lev Michael (eds.), *Amazonian languages: An international handbook*, vol. 1, 65–124. Berlin: De Gruyter Mouton.
- Foley, William A. 1991. *The Yimas language of New Guinea*. Stanford: Stanford University Press.
- Fortescue, Michael. 2007. *Comparative Wakashan dictionary*. Munich: LINCOM.
- Fortuin, Egbert. 2023. Unbounded repetition, habituality, and aspect from a comparative perspective. *Folia Linguistica* 57(1). 135–175. <https://doi.org/10.1515/flin-2022-2054>

- François, Alexandre. 2021. Teanu dictionary (Solomon Islands). *Dictionaria* 15. 1–1877. <https://doi.org/10.5281/zenodo.5653063>
- Frostad, Benedicte Haraldstad. 2013. *A grammar of Ughele*. Nijmegen: Radboud Universiteit. (Doctoral dissertation.)
- Gamble, Geoffrey. 1978. *Wikchamni grammar*. Berkeley: University of California Press.
- Genis, René & Paula Kyselica. 2026. Habitual constructions in Slovak. In Sune Gregersen & Kees Hengeveld (eds.), *A layered approach to habitual constructions*, 309–338. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.136.10gen>
- Georg, Stefan. 2007. *A descriptive grammar of Ket (Yenisei-Ostyak)*, vol. 1. Folkestone: Global Oriental.
- Givón, T. 2011. *Ute reference grammar*. Amsterdam: John Benjamins.
- Gnanadesikan, Amalia E. 2017. *Dhivehi: The language of the Maldives*. Berlin: De Gruyter Mouton.
- Göksel, Aslı & Celia Kerslake. 2005. *Turkish*. London: Routledge.
- Gold, Elaine. 1999. *Aspect, tense and the lexicon: Expression of time in Yiddish*. Toronto: University of Toronto. (Doctoral dissertation.)
- González, Hebe Alicia. 2005. *A grammar of Tapiete (Tupi-Guarani)*. Pittsburgh: University of Pittsburgh. (Doctoral dissertation.)
- Graczyk, Randolph. 2007. *A grammar of Crow*. Lincoln: University of Nebraska Press.
- Grambank Consortium. 2023. *Grambank*. <https://grambank.clld.org/>
- Gravelle, Gloria. 2010. *A grammar of Moskona: An East Bird's Head language of West Papua, Indonesia*. Amsterdam: VU Amsterdam. (Doctoral dissertation.)
- Green, Thomas M. 1999. *A lexicographic study of Ulwa*. Cambridge, MA: Massachusetts Institute of Technology. (Doctoral dissertation.)
- Gregersen, Sune & Kees Hengeveld (eds.). 2026. *A layered approach to habitual constructions*. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.136>
- Gregersen, Sune, Nils Karsten & Marieke Olthof. 2021. Habituals in contrast: Danish *pleje* and its Dutch and German translations. *Linguistics in Amsterdam* 14(1). 39–64.
- Gregersen, Sune & Eva van Lier. forthc. Habitual serial verb constructions: a cross-linguistic survey. In Roland Pfau, Beyza Sümer & Cornelia Loos (eds.), *Serial predicates across modalities*. Berlin: Language Science Press.
- Grimm, Nadine. 2021. *A grammar of Gyeli*. Berlin: Language Science Press.
- Gunnink, Hilde. 2022. *A grammar of Fwe*. Berlin: Language Science Press.

- Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2024. *Glottolog*, v. 5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://glottolog.org>
- Hancock, Ian F. 1971. *A study of the sources and development of the lexicon of Sierra Leone Krio*. London: SOAS. (Doctoral dissertation.)
- Hancock, Ian F. 1994. Componentiality and the creole matrix: The Southwest English contribution. In Michael Montgomery (ed.), *The crucible of Carolina: Essays in the development of Gullah language and culture*, 95–114. Athens: The University of Georgia Press.
- Hannß, Katja. 2008. *Uchumataqu: The lost language of the Urus of Bolivia*. Nijmegen: Radboud University. (Doctoral dissertation.)
- Harms, Philip L. 1994. *Epena Pedee syntax*. Dallas: Summer Institute of Linguistics.
- Harrington, J. P. 1928. Vocabulary of the Kiowa language. *Bureau of American Ethnology Bulletin* 84. 1–255. <http://hdl.handle.net/10088/15551>
- Haspelmath, Martin. 1993. *A grammar of Lezgian*. Berlin: Mouton de Gruyter.
- Haspelmath, Martin. 1998. The semantic development of old presents: New futures and subjunctives without grammaticalization. *Diachronica* 15(1). 29–62. <https://doi.org/10.1075/dia.15.1.03has>
- Hatfield, Adam. 2016. *A grammar of Mehek*. Buffalo: SUNY Buffalo. (Doctoral dissertation.)
- Haude, Katharina. 2006. *A grammar of Movima*. Nijmegen: Radboud University. (Doctoral dissertation.)
- Hellman, Matias. 2005. *The verbs znati and um(j)eti in Serbian, Croatian and Bosnian: A case study in the grammaticalisation of habitual auxiliaries*. Helsinki: Helsinki University Press.
- Hellwig, Birgit. 2011. *A grammar of Goemai*. Berlin: De Gruyter Mouton.
- Hengeveld, Kees & Rafael Fischer. 2018. A'ingae (Cofán/Kofán) operators. *Open Linguistics* 4. 328–355. <https://doi.org/10.1515/opli-2018-0018>
- Hernández Dominguez, Rolando. 2015. *Na-kan-ne jucha anapu jimpo: Aspecto, tiempo, modalidad y modo en purépecha*. Mexico City: CIESAS. (Doctoral dissertation.)
- Hewitt, B. G. 1979. *Abkhaz*. Amsterdam: North-Holland.
- Hewitt, George. 1995. *Georgian: A structural reference grammar*. Amsterdam: John Benjamins.

- Hill, Archibald A. 1975. The habitative aspect of verbs in Black English, Irish English, and Standard English. *American Speech* 50(3–4). 323–324. <http://www.jstor.org/stable/3088018>
- Hinds, John. 1986. *Japanese*. Beckenham: Croom Helm.
- Horie, Kaoru. 2008. The grammaticalization of nominalizers in Japanese and Korean. In María José López-Couso & Elena Seoane (eds.), *Rethinking grammaticalization: New perspectives*, 169–187. Amsterdam: John Benjamins.
- Hualde, José I. & Jon Ortiz de Urbana. 2003. *A grammar of Basque*. Berlin: Mouton de Gruyter.
- Inagaki, Kazuya. 2013. TAME indicators in Kadorih. *NUSA* 55. 95–121. <http://hdl.handle.net/10108/74328>
- Ingram, Andrew. 2001. *Anamuxra: A language of Madang Province, Papua New Guinea*. Sydney: University of Sydney. (Doctoral dissertation.)
- Ingram, Andrew. 2006. Serial verb constructions in Dumo. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *Serial verb constructions: A cross-linguistic typology*, 202–222. Oxford: Oxford University Press.
- Jacobsen, William H. 1964. *A grammar of the Washo language*. Berkeley: University of California, Berkeley. (Doctoral dissertation.)
- Janhunen, Juha. 2012. *Mongolian*. Amsterdam: John Benjamins.
- Jany, Carmen. 2009. *Chimariko grammar*. Berkeley: University of California Press.
- Jaroenkitboworn, Kandaporn. 2009. The polysemy of “chôp” in Thai: A pragmatically motivated phenomenon. *Manusya: Journal of Humanities* 12(3). 75–82. <https://doi.org/10.1163/26659077-01203006>
- Jędrzejowski, Łukasz. 2021. On the habitual verb *pflügen* in German: Its use, origin, and development. *Linguistics* 59(6). 1473–1530. <https://doi.org/10.1515/ling-2021-0161>
- Jenny, Mathias & San San Hnin Tun. 2016. *Burmese*. London: Routledge.
- Johnson, Heidi Anna. 2000. *A grammar of San Miguel Chimalapa Zoque*. Austin: University of Texas at Austin. (Doctoral dissertation.)
- Kaiser, Stefan, Yasuko Ichikawa, Noriko Kobayashi & Hilofumi Yamamoto. 2013. *Japanese*, 2nd edn. London: Routledge.
- Kapitonov, Ivan. 2021. *A grammar of Kunbarlang*. Berlin: De Gruyter Mouton.
- Kasak, Ryan. 2019. *Affix ordering and templatic morphology in Mandan*. New Haven: Yale University. (Doctoral dissertation.)

- Kennard, Edward. 1936. Mandan grammar. *International Journal of American Linguistics* 9(1). 1–43. <https://doi.org/10.1086/463817>
- Kim, Jong-Bok. 2016. *The syntactic structures of Korean*. Cambridge: Cambridge University Press.
- Kiso, Andrea. 2012. *Tense and aspect in Chichewa, Citumbuka and Cisená*. Stockholm: Stockholm University. (Doctoral dissertation.)
- Klokeid, Terry J. 1976. Encliticization in Nitinaht. *International Conference on Salishan Languages* 11. 211–246.
- Klomp, Ulrika. 2021. *A descriptive grammar of Sign Language of the Netherlands*. Amsterdam: University of Amsterdam. (Doctoral dissertation.)
- Kluge, Angela. 2023. Serial verb constructions in Papuan Malay: Forms, functions, and indeterminacy. *Journal of the Southeast Asian Linguistics Society* 16(1). 1–36. <http://hdl.handle.net/10524/52507>
- Kouwenberg, Silvia. 1993. *A grammar of Berbice Dutch Creole*. Berlin: Mouton de Gruyter.
- Kruspe, Nicole D. 2004. *A grammar of Semelai*. Cambridge: Cambridge University Press.
- Kuteva, Tania, Bernd Heine, Bo Hong, Haiping Long, Heiko Narrog & Seongha Rhee. 2019. *World lexicon of grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.
- Lai, I-Wen. 2009. *Time in the Iquito language*. Austin: University of Texas at Austin. (Doctoral dissertation.)
- Lewis, M. Paul, Gary F. Simons & Charles D. Fennig. 2015. *Ethnologue*, 18th edn. Dallas: SIL International.
- Liljegren, Henrik. 2016. *A grammar of Palula*. Berlin: Language Science Press.
- Linn, Mary S. 2000. *A grammar of Euchee (Yuchi)*. Lawrence: University of Kansas. (Doctoral dissertation.)
- Lipkind, William. 1945. *Winnebago grammar*. New York: King's Crown Press.
- Louwerse, John. 1988. *The morphosyntax of Una in relation to discourse structure*. Canberra: ANU.
- Lovegren, Jesse. 2013. *Mungbam grammar*. Buffalo: SUNY Buffalo. (Doctoral dissertation.)
- MacKay, Carolyn J. 1999. *A grammar of Misanla Totonac*. Salt Lake City: University of Utah Press.
- Marlett, Stephen A. 2016. *Cmiiq Itom: The Seri language*. Unpublished MS.

- Maslova, Elena. 2003. *A grammar of Kolyma Yukaghir*. Berlin: Mouton de Gruyter.
- Mattiola, Simone. 2019. *Typology of pluractional constructions in the languages of the world*. Amsterdam: John Benjamins.
- Mattiola, Simone. 2020. Pluractionality: A cross-linguistic perspective. *Language and Linguistics Compass* 14(3). e12366. <https://doi.org/10.1111/lnc3.12366>
- Mattissen, Johanna. 2003. *Dependent-head synthesis in Nivkh: A contribution to a typology of polysynthesis*. Amsterdam: John Benjamins.
- McKenzie, Andrew. 2021. Two types of habituals: Kiowa ingredients of a modular imperfective. *Sinn und Bedeutung* 25. 636–647. <https://doi.org/10.18148/sub/2021.v25i0.958>
- McPherson, Laura. 2020. *A grammar of Seenku*. Berlin: De Gruyter Mouton.
- McWhorter, John. 2019. Pidgins and creoles. In *The Oxford Research Encyclopedia of Linguistics*. <https://doi.org/10.1093/acrefore/9780199384655.013.261>
- Meakins, Felicity & Patrick McConvell. 2021. *A grammar of Gurindji*. Berlin: De Gruyter Mouton.
- Mead, David & Scott Youngman. 2008. Verb serialisation in Tolaki. In Gunter Senft (ed.), *Serial verb constructions in Austronesian and Papuan languages*, 113–139. Canberra: Pacific Linguistics.
- Mixco, Mauricio. 1997. *Mandan*. Munich: LINCOM.
- Mihas, Martin. 2015. *A grammar of Alto Perené (Arawak)*. Berlin: De Gruyter Mouton.
- Miller, Amy. 2001. *A grammar of Jamul Tiipay*. Berlin: Mouton de Gruyter.
- Miyaoka, Osahito. 2012. *A grammar of Central Alaskan Yupik (CAY)*. Berlin: Mouton de Gruyter.
- Molochieva, Zarina. 2010. *Tense, aspect, and mood in Chechen*. Leipzig: Leipzig University. (Doctoral dissertation.)
- Mulder, Mijke. 2021. A first investigation into the encoding of habitual aspect in Muklom with reference to other Sal languages. *Linguistics in Amsterdam* 14(1). 105–121.
- Myanmar–English Dictionary*. 1993. Yangon: Myanmar Language Commission. <http://sealang.net/burmese/dictionary.htm>
- Nedjalkov, Igor. 1997. *Evenki*. London: Routledge.
- Nedjalkov, Vladimir P. & Galina A. Otaina. 2013. *A syntax of the Nivkh language: The Amur dialect*. Amsterdam: John Benjamins.

- Neels, Jakob. 2015. The history of the quasi-auxiliary *use(d) to*: A usage-based account. *Journal of Historical Linguistics* 5(2). 177–234. <https://doi.org/10.1075/jhl.5.2.01nee>
- Nikolaeva, Irina. 2014. *A grammar of Tundra Nenets*. Berlin: De Gruyter Mouton.
- Nordlinger, Rachel. 1998. *A grammar of Wambaya, Northern Territory (Australia)*. Canberra: ANU.
- Nurse, Derek. 2008. *Tense and aspect in Bantu*. Oxford: Oxford University Press.
- O'Brien, Colleen Alena. 2018. *A grammatical description of Kamsá, a language isolate of Colombia*. Honolulu: University of Hawai'i at Mānoa. (Doctoral dissertation.)
- O'Donovan, John. 1845. *A grammar of the Irish language*. Dublin: Hodges & Smith.
- Olawsky, Knut. 2006. *A grammar of Urarina*. Berlin: Mouton de Gruyter.
- Olsson, Bruno. 2021. *A grammar of Coastal Marind*. Berlin: De Gruyter Mouton.
- Omamor, Augusta Phil. 1982. Tense and aspect in Işekiri. *Journal of West African Languages* 12. 95–129.
- Parks, Douglas R. 1976. *A grammar of Pawnee*. New York: Garland.
- Paudyal, Netra Prasad. 2015. *Aspects of Chintang syntax*. Zurich: University of Zurich. (Doctoral dissertation.)
- Payne, Doris L. & Thomas E. Payne. 1990. Yagua. In Desmond C. Derbyshire & Geoffrey K. Pullum (eds.), *Handbook of Amazonian languages*, vol. 2, 249–474. Berlin: Mouton de Gruyter.
- Payne, Thomas E. & Doris L. Payne. 2012. *A typological grammar of Panare: A Cariban language of Venezuela*. Leiden: Brill.
- Perri Ferreira, Helder. 2017. *Yanomama clause structure*. Nijmegen: Radboud University. (Doctoral dissertation.)
- Quesada, J. Diego. 2000. *A grammar of Teribe*. Munich: LINCOM.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
- Redden, James E. 1990. Serial vs. consecutive verbs in Walapai. In Brian D. Joseph & Arnold M. Zwicky (eds.), *When verbs collide: Papers from the 1990 Ohio State Mini-Conference on Serial Verbs*, 240–246. Columbus: Ohio State University. <http://hdl.handle.net/1811/82129>
- Reintges, Chris H. 2004. *Coptic Egyptian (Sahidic dialect)*. Cologne: Köppe.
- Remijsen, Bert. 2010. Nouns and verbs in Magey Matbat. In M. C. Ewing & Marian Klamer (eds.), *Typological and areal analyses: Contributions from East Nusantara*, 285–316. Canberra: Pacific Linguistics.

- Renck, G. L. 1975. *A grammar of Yagaria*. Canberra: ANU.
- Rice, Keren. 1989. *A grammar of Slave*. Berlin: Mouton de Gruyter.
- Rickford, John R. 1974. The insights of the mesolect. In David DeCamp & Ian F. Hancock (eds.), *Pidgins and creoles: Current trends and prospects*, 92–117. Washington, D.C.: Georgetown University Press.
- Rigden, Veda. No date. *Karkar grammar essentials*. Unpublished MS.
- Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314. <https://doi.org/10.1515/lity.1998.2.3.263>
- Ring, Hiram. 2015. *A grammar of Pnar*. Singapore: Nanyang Technological University. (Doctoral dissertation.)
- Romeo, Nicoletta. 2008. *Aspect in Burmese: Meaning and function*. Amsterdam: John Benjamins.
- Romero-Figeroa, Andrés. 1997. *A reference grammar of Warao*. München: LINCOM.
- Round, Erich R. 2009. *Kayardild morphology, phonology and morphosyntax*. New Haven: Yale University. (Doctoral dissertation.)
- Sakel, Jeanette. 2004. *A grammar of Mosestén*. Berlin: Mouton de Gruyter.
- Salminen, Mikko. 2017. *A grammar of Umbeyajts as spoken by the Ikojts people of San Dionisio del Mar, Oaxaca, Mexico*. Douglas: James Cook University. (Doctoral dissertation.)
- Sarvasy, Hannah S. 2017. *A grammar of Nungon: A Papuan language of Northeast New Guinea*. Leiden: Brill.
- Schwegler, Armin. 2013. Palenquero structure dataset. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *The Atlas of Pidgin and Creole Language Structures online*. <http://apics-online.info/contributions/48>
- Shimelman, Aviva. 2017. *A grammar of Yauyos Quechua*. Berlin: Language Science Press.
- Shipley, William F. 1963. *Maidu texts and dictionary*. Berkeley: University of California Press.
- Shipley, William F. 1964. *Maidu grammar*. Berkeley: University of California Press.
- Silverstein, Michael. 1974. Dialectal developments in Chinookan tense-aspect systems: An areal-historical analysis. *International Journal of American Linguistics* 40(4). S45–S99. <http://www.jstor.org/stable/42004710>
- Singerman, Adam. 2018. *The morphosyntax of Tuparí, a Tupían language of the Brazilian Amazon*. Chicago: The University of Chicago. (Doctoral dissertation.)

- Smeets, Ineke. 2008. *A grammar of Mapuche*. Berlin: Mouton de Gruyter.
- Sohn, Sung-Ock S. 1995. *Tense and aspect in Korean*. Honolulu: University of Hawai'i Press.
- Sonnenschein, Aron Huey. 2004. *A descriptive grammar of San Bartolomé Zoogocho Zapotec*. Los Angeles: University of Southern California. (Doctoral dissertation.)
- Spencer, Katharine. 2008. Kwomtari grammar essentials. In Murray Honsberger, Carol Honsberger & Ian Tupper (eds.), *Kwomtari phonology and grammar essentials*, 53–180. Ukarumpa: SIL-PNG.
- Sposato, Adam. 2021. *A grammar of Xong*. Berlin: De Gruyter Mouton.
- Stahlke, Herbert. 1970. Serial verbs. *Studies in African Linguistics* 1(1). 60–99.
- Stapert, Eugénie. 2026. The habitual construction in Dolgan. In Sune Gregersen & Kees Hengeveld (eds.), *A layered approach to habitual constructions*, 121–144. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.136.04sta>
- Stenzel, Kristine. 2013. *A reference grammar of Kotiria (Wanano)*. Lincoln: University of Nebraska Press.
- Strange, David. 1973. Indicative and subjunctive in Upper Asaro. *Linguistics* 11(110). 82–97. <https://doi.org/10.1515/ling.1973.11.110.82>
- Suttles, Wayne. 2004. *Musqueam reference grammar*. Vancouver: UBC Press.
- Svantesson, Jan-Olof & Arthur Holmer. 2014. Kammu. In Mathias Jenny & Paul Sidwell (eds.), *The handbook of Austroasiatic languages*, vol. 2, 957–1002. Leiden: Brill.
- Taylor, Gerald. 2000. *Estudios lingüísticos sobre Chachapoyas*. Lima: Instituto Francés de Estudios Andinos.
- Terhart, Lena. 2024. *A grammar of Paunaka*. Berlin: Language Science Press. <https://doi.org/10.5281/zenodo.10517531>
- Terrill, Angela. 2003. *A grammar of Lavukaleve*. Berlin: Mouton de Gruyter.
- Thieroff, Rolf. 2000. On the areal distribution of tense-aspect categories in Europe. In Östen Dahl (ed.), *Tense and aspect in the languages of Europe*, 265–308. Berlin: Mouton de Gruyter.
- Thiesen, Wesley & David Weber. 2012. *A grammar of Bora with special attention to tone*. Dallas: SIL International.
- Thornes, Tim. 2003. *A Northern Paiute grammar with texts*. Eugene: University of Oregon. (Doctoral dissertation).

- Thornes, Tim. 2011. Dimensions of Northern Paiute multi-verb constructions. In Alexandra Y. Aikhenvald & Pieter Muysken (eds.), *Multi-verb constructions: A view from the Americas*, 27–61. Leiden: Brill.
- Tsunoda, Tasaku. 2020. Modern Standard Japanese. In Tasaku Tsunoda (ed.), *Mermaid construction: A compound-predicate construction with biclausal appearance*, 65–124. Berlin: De Gruyter Mouton.
- Ultan, Russell. 1967. *Konkow grammar*. Berkeley: UC Berkeley. (Doctoral dissertation.)
- Vajda, Edward. 2004. *Ket*. Munich: LINCOM.
- Vajda, Edward. 2020. Areal features in Yeniseian grammaticalization. In Walter Bisang & Andrej Malchukov (eds.), *Grammaticalization scenarios from Europe and Asia*, 433–464. Berlin: De Gruyter Mouton.
- Valenzuela, Pilar M. 2003. *Transitivity in Shipibo-Konibo grammar*. Eugene: University of Oregon. (Doctoral dissertation.)
- Vallejos, Rosa. 2016. *A grammar of Kukama-Kukamiria: A language from the Amazon*. Leiden: Brill.
- van Boven, Cindy & Marloes Oomen. 2021. Habituals in Sign Language of the Netherlands: A corpus-based study. *Linguistics in Amsterdam* 14(1). 160–184.
- van den Berg, René. 1989. *A grammar of the Muna language*. Dordrecht: Foris.
- van den Berg, René. 2019. Compounds and serial verbs in Vitu: A scalar approach. *Language and Linguistics in Oceania* 11. 1–18.
- van der Leeden, Alexander C. No date. *Ma'ya grammar*. Unpublished manuscript.
- van der Vliet, Jacques & Ewa D. Zakrzewska. 2026. The habitual construction in Sahidic Coptic. In Sune Gregersen & Kees Hengeveld (eds.), *A layered approach to habitual constructions*, 42–79. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.136.02van>
- van der Voort, Hein. 2003. Reduplication of person markers in Kwaza. *Acta Linguistica Hafniensia* 35(1). 65–94. <https://doi.org/10.1080/03740463.2003.10416074>
- van der Voort, Hein. 2004. *A grammar of Kwaza*. Berlin: Mouton de Gruyter.
- van der Voort, Hein. 2026. Habitual constructions in Kwaza. In Sune Gregersen & Kees Hengeveld (eds.), *A layered approach to habitual constructions*, 170–201. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.136.06van>
- van Enk, Gerrit J. & Lourens de Vries. 1997. *The Korowai of Irian Jaya: Their language in its cultural context*. Oxford: Oxford University Press.
- van Gijn, Erik. 2006. *A grammar of Yurakaré*. Nijmegen: Radboud University. (Doctoral dissertation.)

- van Schaaik, Gerjan. 2020. *The Oxford Turkish grammar*. Oxford: Oxford University Press.
- van Staden, Miriam. 2000. *Tidore: A linguistic description of a language of the North Moluccas*. Leiden: Leiden University. (Doctoral dissertation.)
- Vázquez Álvarez, Juan Jesús. 2011. *A grammar of Chol, a Mayan language*. Austin: University of Texas at Austin. (Doctoral dissertation.)
- Verhaard, John W. M. 1995. *Toward a reference grammar of Tok Pisin*. Honolulu: University of Hawai'i Press.
- Vidal, Alejandra. 2001. *Pilagá grammar*. Eugene: University of Oregon. (Doctoral dissertation.)
- Walker, Neil Alexander. 2013. *A grammar of Southern Pomo: An indigenous language of California*. Santa Barbara: UC Santa Barbara. (Doctoral dissertation.)
- Wang, Peter, Robert Hunt, Jeff McGriff & Richard E. Elkins. 2006. *The grammar of Matigsalug Manobo*. Quezon City: SIL Philippines.
- Watkins, Laurel J. 1984. *A grammar of Kiowa*. Lincoln: University of Nebraska Press.
- Werner, Heinrich. 2002. *Vergleichendes Wörterbuch der Jenissej-Sprachen*, vol. 1. Wiesbaden: Harrassowitz.
- Widmer, Manuel. 2017. *A grammar of Bunan*. Berlin: De Gruyter Mouton.
- Wilson, Jennifer. 2017. *A grammar of Yeri, a Torricelli language of Papua New Guinea*. Buffalo: SUNY Buffalo. (Doctoral dissertation.)
- Wolvengrey, Arok. 2026. Habitual constructions in *nēhiyawēwin* (Plains Cree). In Sune Gregersen & Kees Hengeveld (eds.), *A layered approach to habitual constructions*, 80–120. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.136.03wol>
- Wurm, S. A. 1977. Missionary lingue franche: Kiwai. In S. A. Wurm (ed.), *New Guinea area languages and language study*, vol. 3, 893–906. Canberra: ANU.
- Xrakovskij, Viktor S. 1997a. Semantic types of the plurality of situations and their natural classification. In Viktor S. Xrakovskij (ed.), *Typology of iterative constructions*, 3–64. Munich: LINCOM.
- Xrakovskij, Viktor S. (ed.). 1997b. *Typology of iterative constructions*. Munich: LINCOM.
- Yillah, Sorie M. & Chris Corcoran. 2007. Krio (Creole English). In John Holm & Peter L. Patrick (eds.), *Comparative creole syntax*, 175–198. London: Battlebridge.
- Yoder, Brendon E. 2020. *A grammar of Abawiri, a Lakes Plain language of Papua, Indonesia*. Santa Barbara: UC Santa Barbara. (Doctoral dissertation.)

- Yoshioka, Noboru. 2012. *A reference grammar of Eastern Burushaski*. Tokyo: Tokyo University of Foreign Studies. (Doctoral dissertation.)
- Young, R. A. 1971. *The verb in Bena-bena: Its form and function*. Canberra: Pacific Linguistics.
- Zahrer, Alexander. 2023. *A grammar of Muyu*. Münster: University of Münster. (Doctoral dissertation.)
- Zamponi, Raoul & Bernard Comrie. 2021. *A grammar of Akajeru*. London: UCL Press.

CONTACT

sngs@proton.me

e.h.vanlier@uva.nl

From construction to suffix: Grammaticalization and areal diffusion of jussives and optatives based on ‘say’ in Daghestan

NINA DOBRUSHINA

CNRS/UNIVERSITÉ LYON 2

Submitted: 21/05/2025 Revised version: 15/09/2025

Accepted: 02/10/2025 Published: 24/07/2026



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

Abstract

In Nakh-Daghestanian languages (Northern Caucasus), one of the most common sources of the jussive (indirect command such as ‘Let him go’) are grammaticalized from reported speech constructions with the imperative of the main verb followed by the imperative of the verb ‘say’: e.g., literally *go.IMP say.IMP*. In some languages, the imperative of the verb ‘say’ is reduced to a suffix. Suffixes coming back to the verb ‘say’ can also be found in optatives (forms that express the wish of the speaker, such as ‘May God bless you’). The aim of this paper is to model the process of grammaticalization of reported speech constructions into jussive or optative, exploring the accompanying morphosyntactic restructuring and semantic changes in a dozen of languages of the Nakh-Daghestanian family. The paper also discusses possible scenarios of the diffusion of the forms with the ‘say’-element spanning the whole territory where languages from different branches of the family are spoken.

Keywords: Jussives; optatives; grammaticalization; reported speech; Nakh-Daghestanian family; diffusion

1. Introduction

Most languages have highly grammaticalized ways of expressing an indirect command (*third person imperative*, aka *jussive*, Gusev 2013: 210). The wish of the

speaker (*optative*) may also have a grammaticalized expression. In English (eng; Indo-European, Germanic),¹ jussives are typically encoded by the verb ‘let’ (1), whereas optatives can be encoded by the modal verb ‘may’ (2):

(1) *If he needs money, let him earn it!*

(2) *May you have a long and fruitful marriage.*

Optatives and jussives are rarely expressed by synthetic forms in modern Indo-European languages (but see Rusakov & Morozova 2022 about the Albanian optative). Conversely, in Nakh-Daghestanian (East Caucasian) languages, forms of jussives and optatives can be found at different stages of grammaticalization – from the constructions with auxiliary verbs or participles to synthetic forms with bound affixes. A frequent way to encode indirect commands in Nakh-Daghestanian languages is to use the suffix that is based on the verb ‘say’ (see Dobrushina 2012, Authier 2015, Dobrushina 2025):

(3) Chamalal (cji; Nakh-Daghestanian, Avar-Andic-Tsezic; Magomedova 2004: 45; quoted from Mauri & Sansò 2011: 3503)

a. *w-un-abe-ǰ’a*
1-go-IMP-OPT
‘Let him go!’

b. *ǰ’a*
speak-IMP
‘Say!’

As I will demonstrate in this paper, the ‘say’-pattern for jussives and optatives has spread across a large territory with a dozen Nakh-Daghestanian languages, including languages belonging to different branches. Cross-linguistically, however, grammaticalization of the verb ‘say’ to jussives seems to be a rare phenomenon. It is not mentioned in the studies of reported speech in Africa (Güldemann 2008), extended usages of the verb ‘say’ in Siberia (Matić & Pakendorf 2013), and reported

¹ Here and in what follows, when naming a language for the first time, I provide its ISO 639-3 code, as well as its genetic classification according to Glottolog.

speech constructions cross-linguistically (Spronck & Casartelli 2021); nor is it present in the lexicon of grammaticalization by Kuteva et al. (2019). So far, I am aware of three such languages outside the Nakh-Daghestanian family: Dongolese Nubian (kzh; Nubian), Armbruster 1960: 391, Burmese (mya; Sino-Tibetan, Burmo-Qiangic, Bernot 1980: 161–162, via Gusev 2013: 212), and Judeo-Tat (jdt; Indo-European, Indo-Iranian, Authier 2012: 183); the latter is spoken in Daghestan.

In order to understand the mechanism of the remarkable spread of this rare grammaticalization path in Daghestan, I will consider the development of jussives and optatives with the ‘say’-element in thirteen Nakh-Daghestanian languages. I will show that jussive and optative forms originate in the reported speech construction with two imperative clauses: *V-IMP- ‘say’-IMP*. I will trace the entire path from this construction to the suffix in detail and demonstrate the morphosyntactic and semantic changes that appear at different stages of grammaticalization. I will also discuss possible scenarios for the diffusion of forms with the ‘say’-element.

I will start by presenting the categories in the study, defining the terms, and introducing Nakh-Daghestanian languages and Nakh-Daghestanian jussives and optatives (Section 2). Firstly, the patterns of grammaticalization of the verb ‘say’ by periphrastic jussive constructions in Mehweb (dar; Nakh-Daghestanian, Dargwa) will be examined. I will then present other languages with ‘say’-based jussives and ‘say’-based optatives, analyze some structural properties of these forms, and suggest the grammaticalization path from the verb ‘say’ to the optative suffix as suggested by the data from Nakh-Daghestanian languages (Section 3). I will consider scenarios of the diffusion of ‘say’-jussives and optatives in Daghestan in Section 4, and discuss the reasons that have contributed to the spread of the ‘say’-pattern across this large territory in the conclusion (Section 5).

2. Jussives and optatives

There is no consensus on the use of the terms *jussive* and *optative*. One of the reasons for this is that they are rarely expressed by inflectional forms, especially in Europe. The only typological study of optatives (Dobrushina 2013) discovered 48 languages with a morphologically dedicated optative in the sample of 319 languages, and most were found in one of two areas, the Caucasus and Tibet. The actual ratio of languages with dedicated optatives should be much lower, because the sample was not balanced – the chapter on optatives took into consideration as many languages of the Caucasus

as possible. Inflectional jussives seem to be more widely spread, but I am unaware of any typological studies of their distribution.

2.1. Defining jussives and optatives

In this paper, the term *jussive* is used for the forms or constructions that express commands directed towards someone who is not a participant of the speech act, while the interlocutor serves as a mediator between the speaker and this participant (Dobrushina 2012), such as the French (fra; Indo-European) *Qu'il vienne chez moi!*. This category is sometimes called the third person *imperative* (Xrakovskij 2001; Aikhenvald 2010), *exhortative* (Ammann & van der Auwera 2004), or *indirect command* (Aikhenvald 2010: 210).

The term *optative* will be used for the forms or constructions that express a speaker's wish or a hope that something would happen, cf. English *If only I had money*, or *May you live long and happily* (Bybee et al. 1994: 321, Nikolaeva 2016: 77).

Jussives and optatives are semantically rather close, since both categories imply the wish of the speaker. They are often encoded by the same form or construction. For example, the English construction with *let* is more typical of indirect commands (4), but occasionally it may also be found in the expressions of wish (5):

(4) *Let them all come!*

(5) *Let him be happy without me!*

The important distinction between jussives and optatives is the controllability of the situation. As was pointed out in Dobrushina (2013), the state of affairs wished for is typically outside the speaker's sphere of influence in the case of optatives. With jussives, the speaker has more control of the situation for two reasons. Firstly, there is a person responsible for transmitting the command (the hearer) to the person who will fulfil it; secondly, it is in the power of the latter to implement it.

The terms *jussive* and *optative* can be applied to dedicated morphological forms (Dobrushina 2011, 2013) or used for constructions that express the wish of the speaker but do not contain lexical items, the primary meaning of which is that of wish or desire (Grosz 2012: 5). Both cases are encountered in Nakh-Daghestanian languages.

2.2. Jussives and optatives in Nakh-Daghestanian languages

Nakh-Daghestanian (East Caucasian) languages are a family with a time depth comparable to that of Indo-European languages, spanning a relatively small territory in the highlands of the Northern Caucasus. They are spoken in the republics of Daghestan, Chechnya, and Ingushetia (Russia), with spillovers in Northern Azerbaijan and Eastern Georgia. Depending on the researcher, there are 40 to 50 languages belonging to several distinct branches (Nichols 2003, Koryakov 2006), see Figure 1.

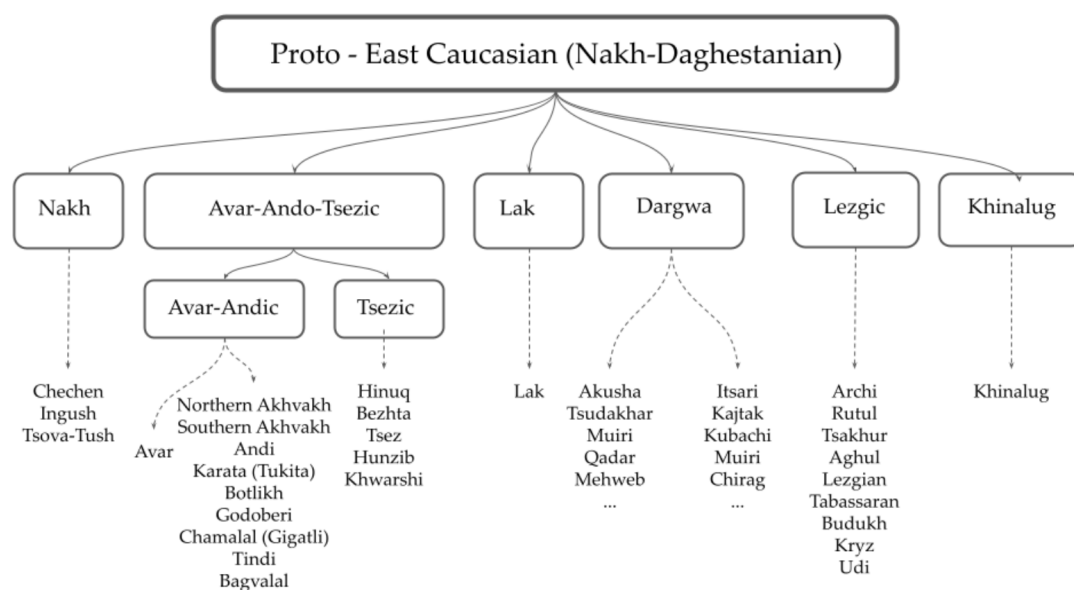


Figure 1. Nakh-Daghestanian family (courtesy of Michael Daniel).

Nakh-Daghestanian languages are morphologically ergative, which means that the sole argument (S) of intransitive verbs is grouped together with P of transitive verbs, but separately from A of transitive verbs: S and P arguments bear nominative (or, in a different terminology, absolutive) case, while A arguments bear the ergative case. Most languages have the category of gender distinguishing from two to six values. Verbs often have slots for prefixal gender agreement. The agreement is usually with the nominative (S/P) argument.

Nakh-Daghestanian languages are special in that they tend to have a dedicated inflectional form for optatives. According to a recent survey, out of 37 Nakh-Daghestanian languages, only one does not have a dedicated optative (Dobrushina 2021). According to another survey, dedicated jussives are much less frequent: only 7 languages out of 37 have them (Dobrushina 2025). This does not signify that the

meaning of indirect command is not expressed in the other 30 languages; there are several other ways to convey it.

The most frequent way is to co-express optative and jussive meanings in one form. In Nakh-Daghestanian languages, such a form is usually available to all persons. It denotes wish (blessings or curses) in the 1/2/3 persons, and command in the 3rd person only. I classify this form as *optative*, reserving the term *jussive* for the form which is available in the third person only (following the terminology used in Dobrushina 2012). Rutul (rut; Nakh-Daghestanian, Lezgetic) instantiates this person/function combination, very typical in Nakh-Daghestanian languages (cf. Table 1).

(6) Rutul (Nakh-Daghestanian, Lezgetic; dialect of Kina, author's field data)

- a. *zi / wi / had* *sax* *r-iš-ij*
 I/ you/ DEM healthy **2-become-OPT**
 'May I (female)/you (female)/she be healthy.'
- b. *said-a* *uq'* *se-j-ij*
 Said-ERG grass **PV-4.mow-OPT**
 'Let Said mow the grass.'
- c. **wa-d* *uq'* *se-j-ij*
 you-ERG grass **PV-4.mow-OPT**
 'You should mow the grass.'

Person	Wish	Command
1		
2		
3		

Table 1. Distribution of the two functions across persons in Rutul optative.

There are, however, Nakh-Daghestanian languages that do not combine the meanings of an indirect command and a wish in the same form. Such languages have a special form or, more often, a construction to encode an indirect command, while the

optative is expressed by a separate form. For example, Chechen (che; Nakh-Daghestanian, Nakh) has several optative forms, and none of these forms are used to encode indirect commands. The form with jussive meaning is built in Chechen as an imperative from the causative of the main verb:

(7) Chechen (Nakh-Daghestanian, Nakh; Zarina Molochieva, p.c.; cited from Dobrushina 2011: 124)

a. third person imperative

p'et'mat-ie bepig d-a-it-a
patimat-ALL bread 3-make-CAUS-IMP
'Let Patimat make bread!'

b. second person wish

lo-hara ho
die-OPT you.SG
'May you die!'

Therefore, Chechen has separate forms for the expression of indirect (third person) commands (jussive) and 1/2/3 person wishes (optative). Another example of the separate expression of the jussive and the optative is Mehweb (Section 3.1).

Quite frequently, jussives and optatives grammaticalize from the reported speech constructions with two imperatives.

3. Grammaticalization of jussives and optatives from the verb 'say'

According to Arkadiev & Maisak (2018: 137), the verb 'say' is among the most grammaticalizable verbs in Nakh-Daghestanian languages. Although the authors do not mention jussives and optatives, they list several other markers that go back to 'say', the most widespread being the *quotative indexes*, i.e., linguistic expressions that signal the presence of a quote. Forms with quotative indexes (or reportatives) come from the combination of two clauses: the reported clause and the matrix clause with the verb 'say'. For example, the reportative *-er* in Archi (aqc; Nakh-Daghestanian, Lezgitic) comes from the imperfective form *war* of the verb 'say'. The marker *-er* is

suffixed to various tense and aspect forms of the lexical verb (Kibrik 1977, Chumakina 2019).

Unlike quotative markers, ‘say’-based jussives and optatives contain lexical verbs in a specific form, namely in the imperative. In Nakh-Daghestanian languages, imperatives can be used in both indirect and direct speech (see, for example, Lezgian,² Haspelmath 1993: 355), one of the indications that the direct-indirect speech division is not very distinct in this family (if present at all, see Daniel 2015, Sumbatova 2019).

In this section, I will trace the grammaticalization path from the reported speech construction with the verb ‘say’ to the jussive/optative suffix in Nakh-Daghestanian languages. In order to do this, I will consider periphrastic jussive constructions in Mehweb and look at the other languages with ‘say’-based jussive and ‘say’-based optatives.

3.1. Periphrastic construction for the jussive: the case of Mehweb

In Mehweb, the meanings of the jussive and optative are expressed separately. The optative is a synthetic form with the suffix *-ab*. The jussive is a periphrastic form which consists of the imperative of the main verb and the imperative of the verb ‘say’ *bet’a*:³

- (8) *saʷi-jal* *uʔq’-e* *bet’-a* *heʔʷan-i* *ɛiz-be-ču*
 <1>self-EMPH 1.go.PFV-IMP say.PFV-IMP similar-ATTR hair-PL-COMIT
 ‘With this kind of haircut, let him drive on his own.’ (Mehweb Corpus)

As with all Mehweb imperatives, the verb *bet’a* ‘say’ retains the ability to encode the number of the interlocutors. Conversely, the embedded lexical verb does not inflect for number (see Section 3.4 for the discussion of plural marking in ‘say’ jussives and optatives).

- (9) *urš-be-jni* *deč’* *b-aq’-a* *bet’-a*
 boy-PL-ERG song 3-do.PFV-IMP say.PFV-IMP
 ‘Let the boys sing a song.’ (addressing one person)

² lez; Nakh-Daghestanian, Lezgic.

³ Unless specifically indicated, all examples on Mehweb in this paper come from Dobrushina (2019).

- (10) *urš-be-jni deč' b-aq'-a bet'-a-na*
 boy-PL-ERG song 3-do.PFV-IMP say.PFV-IMP-IMP.PL
 'Let the boys sing a song.' (addressing several people)

The jussive construction shows some evidence of grammaticalization.

Firstly, while the addressee of the verb 'say' in Mehweb is regularly marked by the interlative case, as in (11), the participant that will carry out the action (below, the *subject* of the jussive construction) bears the marking determined by the verb, including ergative with transitive verbs (12), nominative with intransitive verbs (13), experiential marking with experiential verbs, e.g., dative in Mehweb in (14):

- (11) *musa-ze uz-e bet'-a*
musa-INTER(LAT) 1.work.IPFV-IMP say.PFV-IMP
 'Tell Musa to work.'
- (12) *muhammad-ini deč' b-aq'-a bet'-a*
muhammad-ERG song 3-do.PFV-IMP say.PFV-IMP
 'Let Muhammad sing a song.'
- (13) *musa uz-e bet'-a*
Musa 1.work.IPFV-IMP say.PFV-IMP
 'Let Musa work.'
- (14) *rasuj-s madina d-ig-e bet'-a*
rasul-DAT madina F-love-IMP say.PFV-IMP
 'Let Rasul fall in love with Madina.'

The availability of S or A marking in the jussive construction shows that the jussive has developed into a periphrastic form distinct from the complement construction of the verb 'say'.

Secondly, the verb 'say-IMP' has a fixed position in the jussive: it follows the imperative of the main verb, although in general the word order is not rigid in Nakh–Daghestanian languages. The following sentence is ungrammatical:

- (15) **musa* *bet'-a* *uz-e*
 Musa say.PFV-IMP 1.work.IPFV-IMP
 Intended: 'Let Musa work.'

Thirdly, the meaning of the jussive construction is different from the reported speech construction (cf. the notion of non-canonical *say* in Matić & Pakendorf (2013) and extended reported speech in Spronck & Casartelli 2021: 2). The Mehweb jussive is used in third person commands and permissions (such as '*He wants to go home.* - *Yes, let him go*'), but also in the contexts of indifference, often with inanimate subjects (16) and with the verbs which normally are not used in the form of the imperative (17), (18):

- (16) *rurž-e* *bet'-a* *ħarši*
 boil.IPFV-IMP say.PFV-IMP soup
 'Let the soup boil.'

- (17) **g^w-e!*
 see.PFV-IMP
 Intended: 'See!'

- (18) *g^w-e* *bet'-a!*
 see.PFV-IMP say.PFV-IMP
 'Let him see!' (he should intentionally try to see something)

The meanings of permission and indifference are cross-linguistically very typical for jussives, being described for English (Collins 2004), Hup (Epps 2010), Manambu (Aikhenvald 2010: 74), Kumyk (Dobrushina 2012), Agul (Dobrushina 2012), Russian (Dobrushina 2019), and Livonian (Dailidénaitė 2022). The contexts with an inanimate subject show that *bet'a* in jussive construction does not retain its initial meaning of the speech verb, because one does not address requests to inanimate objects such as soup.

Unlike the forms which will be described in Section 3.3, the Mehweb jussive is very clearly distinguished from the optative. Firstly, the jussive is only available in the third person and cannot be addressed to the hearer or the speaker; first and second person pronouns cannot occur in jussive constructions. Secondly, unlike the optative,

the jussive is not used to express wishes. Example (19) is acknowledged to be grammatical, but semantically inappropriate; one of the Mehweb speakers suggested that this sentence can be uttered by an atheist who thinks that God can be forced to perform an action. The correct choice would be to use the optative (20).

- (19) *?aradeš* *ag-a* *bet'-a*
 health give.PFV-IMP tell.IPFV-IMP
 ? 'Let [Allah] make [you] healthy.'

- (20) *aradeš* *g-a-b!*
 health give.PFV-IRR-OPT
 'May [Allah] make [you] healthy!'

When the jussive is used to denote uncontrollable situations, it is interpreted as an expression of indifference or allowance but never as a wish. For example, the following utterance can be made if the speaker does not care about the rain, e.g., because she has already finished her work in the field:

- (21) *d-aq'-a* *bet'-a* *zab*
 NPL-do.PFV-IMP say.PFV-IMP rain
 'Let it rain (I don't care).'

If the speaker wants the rain to fall, she would rather use the form of the optative:

- (22) *d-aq'-a-b* *zab!*
 NPL-do.PFV-IRR-OPT rain
 'May it rain!'

To sum up, the Mehweb jussive is grammaticalized from the combination of the imperative of the verb 'say' and the imperative of the lexical verb. It is conceived as a transfer of a command or request to the non-locutor via the interlocutor (*Tell him "work!"* → *Let him work!*). Although the Mehweb jussive consists of two verbs, it is grammaticalized. In particular, the verb 'say' has lost its ability to have the inter-lative argument and acquired the ability to embed a clause with an inanimate subject.

3.2. Other ‘say’-based jussives

‘Say’-constructions with jussive meanings are also found in Archi (aqc; Nakh-Dahestanian, Lezgetic), Chirag Dargwa (Nakh-Daghestanian, Dargwa), and Avar (ava; Nakh-Daghestanian, Avar).

Archi allows the variation: the subject can be marked either by contallative or by nominative/ergative/dative, depending on the argument marking of the lexical verb.

(23) Archi (aqc; Nakh-Dahestanian, Lezgetic)

- a. *jasqi-t:u-t* *Patimat-li* / *Patimat-li-r-ši* *lo*
 today-ATTR-4 patimat-ERG / patimat-OBL-CONT-ALL child
barha-ba
nurse.IMP-say.IMP
 ‘This day, let Patimat take care of my child.’
- b. *ošob-č-es* *nak’əm-mul* *χa* *ba*
 brother.PL-OBL-DAT key-PL NPL.find.IMP say.IMP
 ‘Let the brothers find the key.’

Contallative marking is inherited from the verb ‘say’; in Archi, the addressee of speech verbs takes the contallative:

- (24) *ba* *to-r* *ʔanna:-r-ši*
 say.IMP DEM-2 woman.OBL-CONT-ALL
 ‘Tell this woman.’

Unlike in Mehweb, the Archi jussive can have plural marking on the lexical imperative.

(25) Archi (Nakh-Dahestanian, Lezgetic; Kibrik 1977: 220)

- teb* *nen-t’u* *eq’en* *b-oci-r* *ba*
 DEM.PL we-LOC.PL.EMPH(INCL) up.to HPL-stand.IMP-IMP.PL say.IMP
 ‘Let them wait for us.’ (lit. ‘Let them stand till us.’)

The availability of the nominative/ergative marking shows that the Archi jussive is on its way to developing into a periphrastic form distinct from the reported speech construction with the verb ‘say’ and the imperative.

In Chirag Dargwa, the imperative can be combined with the word *iga*, the imperative of the verb ‘say’. The main participant takes ergative/nominative marking, unlike reported speech constructions with the verb ‘say’ (cf. 26). In contrast to Mehweb *bet’a*, Chirag Dargwa *iga* ‘say.IMP’ has lost its ability to add a plural suffix when it is a part of a jussive construction.

(26) Chirag Dargwa (Nakh-Daghestanian, Dargwa; Ganenkov 2026)

- a. *it-e* *c:ad-a-d* *b-arq'-a* *ig-a*
 DEM-PL woman-PL-ERG 3-do.PFV-IMP say.PFV-IMP
 ‘Let those women do it.’

- b. *haraj* *gale* *babaj-la,* *babaj-c:ie* *ba^hrtišan*
 INTJ boy mom-GEN mom-INTER traditional.dish
b-arq'-a *ig-a-j*
 3-do.PFV-IMP say.PFV-IMP-POL
 ‘Come, dear son [literally ‘mother’s son’], please ask your grandmother to make bartishan.’

A similar example is found in the description of Avar. The combination of the imperative of the main verb with the imperative of the verb ‘say’ is used to form the jussive. As follows from the example (27), the subject is marked by the ergative case which is licensed by the verb ‘do’.

(27) Avar (Nakh-Daghestanian, Avar; Alekseev et al. 2012:152)

- hab-e* *ab-e* *dos* *ka-b* *ħalt'i*
 do-IMP say-IMP DEM.ERG this-3 job
 ‘Let him do this job.’

A more advanced stage of grammaticalization can be observed in Lak (Ibe; Nakh-Daghestanian, Lak). The Lak jussive consists of the imperative of the lexical verb +

a morpheme that is derived from the imperative of the verb ‘say’. After reduction, the suffix is no longer identical to the regular imperative of ‘say’:⁴

(28) Lak (Nakh-Daghestanian, Lak)

- a. *nasu:* + *uča* → *nasu:-ča*
 ‘go.imp’ + ‘say.imp’ → ‘let him go!’
- b. *Pat’imat-lu-l* *č:at’* *šaši-ča*
 patimat-OBL-ERG bread **bake.IMP-JUSS**
 ‘Let Patimat bake bread.’ (example courtesy of Rosa Eldarova, quoted in Dobrushina 2011: 123)
- c. *zun* *čijkun* *b-ik’u-ča*
 you.PL.DAT however.you.like **3-be.IMP-JUSS**
 ‘Let (it) be as you wish.’ (Abdullaev 2015: 2105)

Lak form in *-ča* can be occasionally used for wishes, only with 3rd person subject; in general, other forms are used to express wishes in Lak.

- (29) *žahannam-ra-w-un* *nasu:-ča*
 hell-OBL.SG-IN-LAT **go.IMP-JUSS**
 ‘Let him go to hell!’ (Abdullaev 2015: 693)

Similar form is attested in Godoberi (gdo; Nakh-Daghestanian, Andic), where the form in *-ž’a* (from *hiž’a* ‘say.IMP’) is used as jussive.

Table 2 presents all the cases I was able to find in Nakh-Daghestanian languages where the verb ‘say’ is used as the source of the jussive.

In all the languages mentioned in this section (Mehweb, Avar, Archi, and Lak), the construction or the form involving ‘say’ is restricted to the third person and is not used (or is infrequently used) for the expression of wishes. According to my definition, these forms are jussives. There is also a group of languages, the forms of which with

⁴ There are however examples of the Lak jussive where the imperative of ‘say’ is preserved in its entirety, see (37).

‘say’-based suffixes include the meaning of wish (blessings and curses), and in which these forms exist in all persons; according to my definition, they are optatives.

Language	Branch	Construction /Form	Pattern	Example	Availability of addressive encoding of the jussive subject
Mehweb	Dargwa	periphrastic construction	IMP + ‘say’.IMP	<i>uz-e bet'-a</i> 1.work.IPFV-IMP say.PFV-IMP	no
Archi	Lezgian	periphrastic construction	IMP + ‘say’.IMP	<i>barha-ba</i> nurse.IMP -say.IMP	yes
Chirag Dargwa	Dargwa	periphrastic construction	IMP + ‘say’.IMP	<i>b-arq'-a ig-a</i> 3-do.PFV-IMP say.PFV-IMP	no
Avar	Avar-Andic-Tsezic	periphrastic construction	IMP + ‘say’.IMP	<i>hab-e ab-e</i> do-IMP say-IMP	no
Lak	Lak	synthetic form	IMP + suffix (< ‘say’.IMP)	<i>nasu:-ča</i> go.IMP-JUSS	no
Godoberi	Andic	synthetic form	IMP + suffix (< ‘say’.IMP)	<i>ik'ja-λ'a</i> give.IMP-JUSS	no

Table 2. Jussives with the verb ‘say’ in Nakh-Daghestanian languages.

3.3. ‘Say’-based optatives

A number of closely related Andic and Tsezic languages (Northern Akhvakh (akv, Nakh-Daghestanian, Andic), Chamalal (Nakh-Daghestanian, Andic), Tindi (tin; Nakh-Daghestanian, Andic), Hinuq (gin; Nakh-Daghestanian, Tsezic), Karata (kpt; Nakh-Daghestanian, Andic), Khwarshi (khv; Nakh-Daghestanian, Tsezic), and Tsez (ddo; Nakh-Daghestanian, Tsezic) have a form of optative which contains a suffix based on the imperative of the verb ‘say’. The usage of this form is very similar in all these languages. It is compatible with all three persons and functionally follows the pattern described in Section 2 (Table 1): it is used for wishes (blessings and curses) in 1/2/3 persons and for commands only in the 3rd person. In all languages, the *subject* of the optative construction bears the marking determined by the verb, showing no traces of the argument marking of the initial verb ‘say’.

(30) Khwarshi (Nakh-Daghestanian, Tsezic; Khalilova 2009: 249-251)

- a. *j-uhō-λo* *do, heresi is-se* *b-us-ło*
 2-die.IMP-OPT I lie say-IPFV.CVB 3-find-COND
 ‘May I die if I lie!’
- b. *žužah-ma-l* *b-ek’lo-λo* *mížo*
 hell-IN-LAT hpl-fall.IMP-OPT you.PL
 ‘May you go to hell!’
- c. *Pat’imat-i* *j-ezo-λo* *bataxu*
 patimat-ERG 5-buy.IMP-OPT bread
 ‘Let Patimat buy bread.’
- d. *b-uxala* *žumru gollu* *b-uso-λo*
 3-long life be.PRS.PTCP 3-find.IMP-OPT
 ‘May he live a long life!’ (about a newborn)

In some other languages of the Avar-Andic-Tsezic branch, there are also optative suffixes of a different (often unknown) origin: Andi has suffix *-do*,⁵ Avar *-gi* and *-ja-CL*, Bagvalal *-la*, Bezhta *-ala*, Botlikh *-ibe*, *-abu*, Hunzib *-ob*, Southern Akhvakh *-la*.

Andic and Tsezic optatives differ from Mehweb, Archi, Chirag, and Avar forms, which were described in the previous section both functionally and formally. In terms of function, they have broader semantics, since they combine jussive and optative meanings and can be used in all persons. In terms of form, they do not contain a full imperative of the verb ‘say’, but just a part of it. The Lak and Godoberi forms with the ‘say’-suffixes only differ from the Andic and Tsezic optatives functionally.

To sum up, there are at least thirteen Nakh-Daghestanian languages where the jussive or optative is expressed by a form containing the ‘say’-element. In Mehweb, Archi, Chirag, and Avar the whole imperative of the verb ‘say’ is used, whereas in the other languages only a fragment of the imperative of ‘say’ remains.

⁵ According to my recent research, the Andic optative is also derived from a combination of the imperative and the verb ‘say’, with the ending ‘-do’ being a remnant of that verb.

Language	Group	Optative	Stem	Say.IMP
Akhvakh	Andic	-ʒ:’a oxa-ʒ:’a ‘may (s/he) give’	imperative	eʒ:’a
Karata	Andic	-ʒ’a beʒda-ʒ’a ‘may (it) stay’	imperative	keʒ’ã
Chamalal	Andic	-ʒ’a guda-ʒ’a ‘may s/he give’	imperative	iʒ’a
Tindi	Andic	-ʒ’a baχa-ʒ’a ‘may s/he take’	imperative	hiʒ’a
Hinuq	Tsezic	-ʒo kekir-ʒo ‘may s/he send’	imperative	eʒo
Khwarshi	Tsezic	-ʒo (-ʒa) depending on the preceding vowel buso-ʒo ‘may s/he find’	imperative	iʒo
Tsez	Tsezic	-oʒ (-ʒ) depending on the preceding phoneme, e.g., cax-oʒ ‘may s/he write’	imperative	eʒo

Table 3. Optatives with the verb ‘say’ in Nakh-Daghestanian languages.

3.4. Plural marking

Jussives and optatives with the ‘say’ element originate from the combination of two imperatives. Both imperatives, lexical and ‘say’, can potentially encode the number of interlocutors. The ability of the imperative to mark the number of interlocutors varies in Nakh-Daghestanian languages (see the overview, Zerzele 2024). Among the languages discussed in this paper, some do not encode it at all (Northern Akhvakh, Avar, Godoberi, Hinuq, Karata proper, Khwarshi, Tsez), some mark the number only in intransitive verbs (Archi, Chamalal, Anchiq dialect of Karata, Tindi), and some encode it in all verbs (Mehweb, Chirag, Lak). The question is whether the latter seven languages mark the number in ‘say’-jussives/optatives as well as in imperatives.

3.4.1 Plural marking on lexical part

Second person imperatives distinguish singular versus plural in Mehweb, Archi, Chirag, Lak, Chamalal, Tindi and Anchiq dialect of Karata. In Mehweb, Archi and

Chirag ‘say’-forms are periphrastic, in the other four languages they are synthetic. Notably, the lexical part of ‘say’-jussive/optative forms tends to retain plural marking both in periphrastic and synthetic forms. Only in Mehweb does the embedded lexical imperative not retain the number marking when it is a part of the jussive/optative form.

In Archi, intransitive imperatives can optionally take the suffix *-r* to mark the plural interlocutor. Jussive construction preserves the option to have the plural marker on the lexical imperative.

(31) Archi (Nakh-Dagestanian, Lezgetic; Bulbul Musaeva, p.c.; Kibrik 1977: 220)

- a. *č’an inž-i u b-erq^əa ba*
 sheep self- <3>_{EMPH} 3-go.IMP say.IMP
 ‘Let the sheep walk on its own.’ (unassisted; can be said when the sheep was sick, but now it needs to start walking)
- b. *gid-ib noλ’-a-ši b-erq^əa-r ba*
 DEM.PL-3 house-IN-ALL HPL-go.IMP-IMP.PL say.IMP
 ‘Let them come home.’

In Lak, the jussive contains a suffix which originates from the imperative of the verb ‘say’: *uča* ‘say.IMP’ becomes *-ča*. The first part of this form inflects for number (Eldarova 1999: 57, Friedman 2020: 227):

(32) Lak (Nakh-Daghestanian, Lak; Roza Eldarova, p.c.)

- a. *nasu:-ča*
 go.IMP-JUSS
 ‘Let him go.’
- b. *nasi:-ča*
 go.IMP.PL-JUSS
 ‘Let them go.’

In the Anchiq dialect of Karata, the imperative inflects for number, and the optative retains this inflection:

(33) Anchiq Karata (Nakh-Daghestanian, Andic; Magomed Gadžimagomedov, p.c.)

- a. *χeriq* *w-ołi-λ'a* *ho-w*
 back 1-go.IMP-OPT DEM-1
 'Let him go.'
- b. *χeriq* *b-ełibi-λ'a* *ho-be*
 back HPL-go.IMP.PL-OPT DEM-HPL
 'Let them go.'

As follows from these examples, number marking is determined by the number of subjects of jussives/optatives. Table 4 presents the examples of all 'say'-jussives/optatives where the lexical part encodes number.

At the same time, in some languages, which can potentially contain plural marking on the lexical part, it is optional or even less felicitous than singular marking. In Chirag Dargwa, the distinction of the number of the interlocutors is obligatory in 2nd person imperatives, but optional in jussives: the imperative of the lexical verb can appear in the plural form (34a) or remain singular with plural subjects (34b).

(34) Chirag Dargwa (Nakh-Daghestanian, Dargwa; Ganenkov 2026)

- a. *u'l-d-irg-a-ja* *ig-a*
 look-LCT-LV.IPF-IMP-PL say.PFV-IMP
 'Let them look!'
- b. *č'a-b-aq'-a* *ig-a*
 hear-3.SG-LV.PFV-IMP say.PFV-IMP
 'Let them hear this (I don't care).'

In Mehweb, my consultants prefer the lexical imperative to be unmarked for the number in the jussive construction, while the plural marking of the imperative outside the jussive construction is obligatory:

(35) Mehweb (Nakh-Daghestanian, Dargwa)

- a. *deč' b-aq'-a*
 song 3-do.PFV-IMP
 'Sing a song!' (addressing one person)
- b. *deč' b-aq'-a-na*
 song 3-do.PFV-IMP-IMP.PL
 'Sing a song!' (addressing several people)
- c. *urš-be-jni deč' b-aq'-a bet'-a*
 boy-PL-ERG song 3-do.PFV-IMP say.PFV-IMP
 'Let the boys sing a song.' (addressing one person)
- d. ?? *urš-be-jni deč' b-aq'-a-na bet'-a*
 boy-PL-ERG song 3-do.PFV-IMP-IMP.PL say.PFV-IMP
 'Let the boys sing a song.' (addressing several people)

Meanwhile, in the construction which is closer formally to reported speech construction in that the addressee of the verb 'say' is marked by the interlative, the native speaker of Mehweb chose plural marking of the embedded imperative:

- (36) *urš-be-ze deč' b-aq'-a-na bet'-a*
 boy-PL-INTER(LAT) song 3-do.PFV-IMP-IMP.PL say.PFV-IMP
 'Tell boys to sing a song/Let boys sing a song.' (addressing one person) (p.c. with Patimat Sharbuzova)

The tendency to lose plural marking of the embedded imperative found in Mehweb and Chirag Dargwa has a functional explanation. While in the second person imperatives, the number marker denotes the number of the interlocutors, the same marker does not refer to the interlocutor when it is transferred to the third person

form; it rather denotes the number of the agents of the desired situation. This might be the reason why plural marking is optional in the Chirag Dargwa jussive and is less favored in Mehweb. Note that a typological study by Gusev (2013: 2014) revealed that cross-linguistically, third-person imperatives less frequently mark number than other moods.

It is important to add that there is no adequate field data about the actual use of plural forms in other languages where plural marking of the lexical imperative was evidenced (Lak, Chamalal, Tindi, and Anchiq dialect of Karata). It might be that they also tend to vary between marking and not marking plurality.

3.4.2 Plural marking on ‘say’

The verb ‘say’ is transitive, therefore it can express the number of interlocutors only in those languages where both transitive and intransitive imperatives inflect for number – Mehweb, Chirag, and Lak. These three languages behave differently in this respect.

Lak has a synthetic jussive form with the verb ‘say’ reduced to the suffix; it did not retain plural marking:

(37) Lak (Nakh-Daghestanian, Lak; examples courtesy of Rosa Eldarova)

- a. *uča*
say.IMP
‘Say!’ (addressing one person)
- b. *uča:ra*
say.IMP.PL
‘Say!’ (addressing several people)
- c. *wi-l ars t:u-č’a-n ux:u-wča*
you.SG.OBL-GEN son I.OBL-APUD-LAT come.IMP-JUSS
‘Let your (singular) son come to me.’ (addressing one person)

- d. *zu-l* *ars* *tʉ-č'a-n* *ux:u-wča*
 you.PL.OBL-GEN son I.OBL-APUD-LAT come.IMP-JUSS
 'Let your (plural) son come to me.' (addressing several people)

Language	Singular	Plural marking on lexical verb	Plural marking on 'say'
Mehweb	<i>b-aq'-a bet'-a</i> 3-do.PFV-IMP say.PFV-IMP 'Let him/her do'	absent	<i>b-aq'-a bet'a-na</i> 3-do.PFV-IMP say.PFV-IMP-IMP.PL 'Let them do'
Archi (own elicitation; Kibrik 1977: 220)	<i>b-erq^ɛ-a ba</i> 3-walk-IMP say.IMP 'Let it go' (of a donkey)	<i>b-erq^ɛ-a-r ba</i> HPL-walk-IMP-IMP.PL say.IMP 'Let them go'	absent
Chirag Dargwa (Ganenkov 2026)	<i>r-uk-aba ig-a</i> 2-take.PFV-IMP1 say.PFV-IMP 'Let him/her/them take me'	<i>r-uk-aba-ja ig-a</i> 2-take.PFV-IMP1-PL say.PFV-IMP 'Let them take me'	absent
Lak (Roza Eldarova, p.c.)	<i>nasu:-ča</i> go.IMP-JUSS 'Let him go'	<i>nasit:-ča</i> go.IMP.PL-JUSS 'Let them go'	absent
Chamalal (Magomedova 1999: 99)	<i>bu-k'-abe-λ'a</i> 3-be-IMP-OPT 'Let it be'	<i>ba-k''abi-λ'a</i> HPL-be-IMP.PL-OPT 'Let them be'	not applicable
Tindi (Magomed Magomedov, p.c.)	<i>w-ič'-iba-λ'a</i> 1-die-IMP-OPT 'May he die'	<i>b-ič'-ibi-λ'a</i> HPL-die-IMP.PL-OPT 'May they die'	not applicable
Anchik dialect of Karata (Magomed Gadžimagomedov, p.c.)	<i>w-ox-u-λ'a</i> 1-come-IMP-OPT 'Let him come'	<i>b-ex''abi-λ'a</i> 3-come-IMP-OPT 'Let them come'	not applicable

Table 4. Plural marking in 'say'-jussives and optatives.

The Chirag Dargwa jussive is periphrastic. The imperative of the verb 'say' *iga* has a plural marker when it is used in the literal imperative meaning, but the auxiliary part of the jussive cannot appear in the plural form. In the example (38), the jussive is

directed to parents (plural addressees), but the verb ‘say’ does not have plural marking.

(38) Chirag Dargwa (Nakh-Daghestanian, Dargwa; Ganenkov 2026)

*nuš:a-la gale ja^s-ma elg-e **ig-a***

you.PL-GEN boy DEM-LOC stay.PFV-IMP **say.PFV-IMP**

‘Let your (plural) son stay here.’ (as in suggesting that the boy stays overnight at our place)

In contrast, the verb ‘say.IMP’ in the Mehweb periphrastic jussive retains its ability to inflect for plural when it becomes part of the jussive:

(39) *musa-ze uz-e bet'-a-na*

Musa-INTER(LAT) 1.work.IPFV-IMP say.PFV-IMP-IMP.PL

‘Let Musa work.’ (addressing several people)

As it follows from the example (39), the plural of the verb ‘say’ in Mehweb jussive encodes the number of interlocutors, not the number of subjects.

To summarize, number marking can be retained or lost on both parts of ‘say’ jussives and optatives. While the plural of the verb ‘say’ denotes the number of interlocutors, the plural of the lexical verb denotes the number of subjects. This shows that some ‘say’-forms have certain biclausal properties, as discussed in the next section.

3.4.2 ‘Say’-jussives and optatives as univerbation

‘Say’-jussives and optatives present an interesting case of what is called phrasal univerbation – an expression condensed into a word (Lehmann 2020: 214).

There are several indications of the morphological boundness of ‘say’-jussives/optatives.

First of all, the ‘say’-element often shows the loss of autonomy. In many languages, the ‘say’-element is a suffix that cannot be used independently. In other languages, where the imperative of the verb ‘say’ is not reduced to a suffix, it has a fixed position after the lexical verb (see example (15)). In Chirag Dargwa, the verb ‘say’ has lost its ability to take the plural suffix, which also speaks for its reduced autonomy.

The second argument in favor of the morphological boundness of ‘say’-jussives/optatives is that the verbs, which normally cannot be used in the imperative, appear in the form of the imperative as a part of the jussive/optative construction. This was illustrated by Mehweb (example (18)).

On the other hand, there is evidence of the syntactic independence of the embedded lexical verb.

Firstly, it can contain a plural marker in the same way as the independent imperative does, even if the ‘say’-part is a suffix (this happened in Lak, Chamalal, Tindi, and Anchiq Karata, see Section 3.4.1). These forms resemble what is sometimes called morphologically bound complementation, i.e., a construction where a matrix predicate and the head of its clausal complement constitute a single word form but retain their syntactic and semantic independence; in particular, both parts of such constructions have their own sets of arguments (see Maisak 2016, Panova 2018). The ‘say’-element in jussives and optatives can be considered a complement-taking predicate where the lexical part and its dependents are its complement clause. The plural marker on the embedded lexical imperative signals the presence of its argument. The second part of the ‘say’-imperative (originally the matrix verb) can also mark the plurality of its argument in some languages. However, as follows from Table 4, there are no cases where both lexical and ‘say’ imperatives mark plurality. The three forms where it potentially could have happened (Chirag, Mehweb, and Lak), each mark plurality only once, either on the lexical imperative (Chirag and Lak), or on the ‘say’-imperative (Mehweb), thus showing the tendency to lose the autonomy of one of the two parts. The analogy to the morphologically bound complementation is thus not complete.

The second indication of syntactic independence of the lexical verb is the fact that its negative counterpart is prohibitive. Most Nakh-Daghestanian languages have special forms for negative imperatives that are unrelated to the imperatives morphologically and employ a special negation marker (Ochs 2021). Prohibitive, like imperative, is a form that is only used independently, as a separate sentence. As well as prohibitives, the only forms where this marker is found are jussives and optatives. This is not always the case; in some Nakh-Daghestanian languages, optatives use indicative negation, and in some others (e.g., Dargwa branch), the optative has a prohibitive prefix, but the stem of the negative optative is different from that of the prohibitive (Dobrushina 2021).

Meanwhile, ‘say’-based jussives and optatives contain an entire prohibitive form, as in the Northern Akhvakh examples presented in Table 5.

	Positive	Negative
Imperative	<i>o-x:-a</i> 3-give-IMP ‘Give!’	<i>o-x:-uba</i> 3-give-PROH ‘Don’t give!’
Optative	<i>o-x:-a-λ’a</i> 3-give-IMP-OPT ‘May (he) give!’	<i>o-x:-uba-λ’a</i> 3-give-PROH-OPT ‘May (he) not give!’

Table 5. Positive and negative imperatives and optatives in Northern Akhvakh (examples courtesy of Indira Abdullaeva).

Finally, the jussive/optative argument inherits its case from the lexical verb: ergative if the lexical verb is transitive, nominative if the lexical verb is intransitive, dative (or equivalent case) if the verb is experiential (see examples 12, 13, 14). The lexical imperative, therefore, retains a certain independence.

At the same time, the very fact that the lexical verb has an overtly expressed argument speaks for the boundness of the jussive/optative form. Normally, imperative clauses cannot contain third-person subjects. Although there are some special types of contexts when this is possible (only in the presence of the interlocutor; see Dobrushina 2019: 124-125), in the regular situation, imperative clauses are subjectless or contain second person pronouns. Meanwhile, jussive/optative clauses commonly have overt third-person subjects whose case is assigned by the embedded imperative. This means that the subject of jussive/optative clauses is induced by the jussive/optative form as a whole rather than by the embedded imperative: the possibility of having an overt third-person subject is due to the fact that jussives/optatives allow it, unlike second-person imperatives.

‘Say’ jussives/optatives add to the list of the Nakh-Daghestanian forms which exhibit the mixed properties of being morphologically bounded and containing two predicates with certain morphosyntactic independence (see verificatives – Maisak 2016, Daniel & Maisak 2014, and reportatives – Chumakina 2019, Arkadjev & Maisak 2018, Kibrik 1977).

Now, the tentative grammaticalization path for these forms can be discussed.

3.6. Grammaticalization path

In all the languages considered above, the jussive/optative form developed from the reported speech construction with two imperatives. The origin of such a form can be traced based not only on its suffix, but also on its stem. The stem corresponds to the first imperative in the construction (that of the lexical verb), the suffix corresponds to the second imperative (that of the verb ‘say’). Negative counterparts of ‘say’-based jussives and optatives were grammaticalized directly from the reported speech construction with prohibitive and ‘say’-imperatives.

Left-branching syntax, typical of Nakh-Daghestanian languages, could facilitate the coalescence of two imperatives. In Nakh-Daghestanian languages, all arguments tend to be placed to the left of the verb, so that, for example, the sentence ‘Say bring the book to school’ might be literally realized as ‘To school the book bring say’, where the two imperatives are juxtaposed (see examples 11 and 36).

Several facts suggest that ‘say’-optatives occur at a later stage of grammaticalization than jussives. Firstly, while there are jussives which contain the full form of the imperative of the verb ‘say’ (in Mehweb, Archi, Chirag Dargwa, and Avar), there are no such optatives – in optatives, the ‘say’-element is always reduced to a suffix. Secondly, jussives with the ‘say’-element may retain the possibility of argument marking with overtly expressed addressees inherited from the verb ‘say’ (as is partly the case in Archi), while optatives with ‘say’-elements do not show any traces of the argument-marking of the verb ‘say’. Thirdly, there are ‘say’-jussives which do not express an optative meaning (such as in Mehweb or Archi), but there are no ‘say’-optatives which do not also express a jussive meaning. This all points to the fact that ‘say’-optatives develop from ‘say’-jussives by the extension of the jussive meaning to other persons and its transformation to the optative.

The resulting optative form demonstrates further loss of the lexical content of the verb ‘say’. While the jussive often implies the presence of the mediator who will transmit the command to the third person, wishes are meant to be fulfilled by a supernatural power (in Daghestan, usually Allah or, rarely, shaitan in the case of a curse), and there is no need for the mediator between the speaker and the supernatural power. This is especially clear with second-person optatives, such as ‘May you go to hell’:

- (40) Khwarshi (Nakh-Daghestanian, Tsezic; Khalilova 2009: 249-251)

žužah-ma-l b-ek'lo-λo mízho
 hell-IN-LAT HPL-fall.IMP-OPT you.PL.ABS
 ‘May you go to hell!’

The original reported speech construction which underlies this expression is ‘to hell fall.IMP say.IMP you’, where ‘you’ is an argument of the verb ‘fall’. The interlocutor to whom the imperative ‘say’ is addressed and the potential subject of this curse coincide (*‘you tell you “you go to hell”’), which does not make any sense.

In addition, ‘say’-optatives have a wider range of functions compared to jussives, since they express both indirect commands and wishes, thereby covering a broader range of volitional meanings. This can be qualified as the process of *desemanticization*, or *semantic bleaching*, accompanied by the contraction of the ‘say’-imperative (*phonological reduction*).

We can suggest the following grammaticalization path for ‘say’-optatives. The starting point is a reportative construction meaning ‘tell him to do smth’ which contains two verbs: the imperative of the verb ‘say’ and the imperative of the lexical verb. In this construction, the addressee is encoded by the case which is normally used in this language for the addressee with the verb ‘say’:

noun.ADDRESSEE verb.IMP say.IMP

This stage is illustrated by examples (11) and (36).

Later, another argument pattern may develop. Instead of marking the subject of the intended situation as an addressee (licensed by ‘say’), the new argument emerges in the form assigned by the lexical verb: ergative with transitive verb, nominative with intransitive verbs, dative or another case with experiential verbs (examples (12)-(14)):

noun.NOM/ERG/EXP verb.IMP say.IMP

At this stage, variation may be present when the encoding of the arguments follows either the pattern of reported speech with an overtly expressed addressee or is assigned by the lexical verb. This case is illustrated by example (23a). Most likely, at this time, the construction develops other usages, which are incompatible with the

lexical ‘say’. For example, the contexts with an inanimate subject cannot be interpreted as a reported speech construction, such as “Let the soup boil”, where the ‘soup’ cannot be an addressee (example (16)):

animate/inanimate noun.NOM/ERG/EXP verb.IMP say.IMP

Later, the imperative of the verb ‘say’ can be reduced and become a bound suffix (this stage is found in Lak and Godoberi, see examples (28b-c)):

noun.NOM/ERG/EXP verb.IMP-JUSS

Finally, at the last stage, the form extends its meaning to the wishes, and starts being used not only with third-person subjects but also with first and second persons. I classify this form as an optative (see examples (30a-d)).

noun.NOM/ERG/EXP verb.IMP-OPT

In Nakh-Daghestanian languages, different stages of grammaticalization of the verb ‘say’ to jussives and optatives were attested: from the periphrastic forms with two imperatives to synthetic forms with a bound suffix originating from ‘say’.

4. Diffusion of ‘say’-jussives and optatives across Nakh-Daghestanian languages

From the 13⁶ Nakh-Daghestanian languages discussed above, nine languages are closely related, all belonging to the same Avar-Andic-Tsezic branch (Avar, Northern Akhvakh, Chamalal, Godoberi, Hinuq, Karata, Khwarshi, Tindi, and Tsez). The other four languages belong to the other three branches: Dargwa (Mehweb and Chirag), Lezgic (Archi), and Lak, which constitutes a separate branch (Lak) (see Figure 2). Since the pattern of ‘say’ jussives and optatives is typologically infrequent, the following question arises: how did all these languages develop the same jussive/optative grammaticalization pattern? Was it due to inheritance or contact?

⁶ There is, however, one more language which is likely to have ‘say’-optative, but needs special discussion (Dobrushina, accepted for publication) – Rutul, a language of the Lezgic branch.



Figure 2. Map of ‘say’-jussives and optatives (courtesy of Yuri Koryakov).

4.1. Avar-Andic-Tsezic languages

Although ‘say’-suffixes in Andic and Tsezic languages look similar, there are significant differences between them. As follows from Table 6, the verb ‘say’ has a non-ejective lateral λ in Tsezic languages, and an ejective lateral consonant λ' in Andic languages, including Northern Akhvakh, where the ejective lateral is geminated. The same distinction holds for the optative suffixes. The correspondence between non-ejective lateral λ , ejective lateral λ' , and geminated ejective lateral $\lambda\lambda'$ is regular in Andic and Tsezic languages; cf. the word ‘meat’ in those Andic and Tsezic languages which have the same root for it (Table 6).

In the closely related Avar, the verb *abize* (‘say’, infinitive) is not a cognate to the Andic and Tsezic forms. Avar has periphrastic jussives based on the imperative of this verb, *abe* ‘say.IMP’. This verb is not found in Andic and Tsezic languages, nor is the verb with a lateral root found in Avar (Mudrak 2019, 2020). It could have been the case that the proto-Avar-Andic-Tsezic root with a lateral consonant was lost in Avar

and replaced by a new verb (however, note that, according to the reconstruction in Nikolaev & Starostin 1994, both roots existed on the Proto-North-Caucasian level, although their meaning could have changed).

Language	Group	‘meat’	‘say’-IMP	Optative suffix
Northern Akhvakh	Andic	<i>riλː’i</i>	<i>eλː’a</i>	<i>-λː’a</i>
Karata	Andic	<i>riλ’i</i>	<i>keλ’ā</i>	<i>-λ’a</i>
Chamalal	Andic	<i>jiλ’i</i>	<i>iλ’a</i>	<i>-λ’a</i>
Godoberi	Andic	<i>riλ’i</i>	<i>hiλ’a</i>	<i>-λ’a</i>
Tindi	Andic	<i>jiλ’i</i>	<i>hiλ’a</i>	<i>-λ’a</i>
Khwarshi	Tsezic	<i>liλ</i>	<i>iλo</i>	<i>-λo</i>
Tsez	Tsezic	<i>reλ</i>	<i>eλo</i>	<i>-oλ/-λ</i>

Table 6. ‘Meat’, ‘say’ and optative suffix in Andic and Tsezic languages (Kibrik & Kodzasov 1990, Filatov & Daniel 2021).

Therefore, Avar, Andic, and Tsezic languages use their own verbs to construct a jussive/optative. This means that there was no borrowing of matter involved in the diffusion of ‘say’-forms. The inheritance of the whole form from the proto-Avar-Andic-Tsezic is also unlikely. Firstly, because the Avar form is very different from other languages. Secondly, because in Chamalal, Tindi, and Anchiq Karata the lexical imperative retains inflection in number and hence is rather “young”. Thirdly, because in Tsezic languages the suffix looks different depending on the imperative suffix in a respective language (*-λo* in Hinuq, *-λo* (*-λa*) in Khwarshi, *-oλ* (*-λ*) in Tsez), which shows that the development of the synthetic form took place in individual languages.

If we reject the option of borrowing or inheriting synthetic form, we should consider the scenarios of the emergence of a structural pattern, which later developed into synthetic forms in each language. Two such scenarios are possible.

According to the first scenario, that of inheritance, the ‘say’-construction was present at the level of proto-Avar-Andic-Tsezic. It was then retained in some Avar-Andic-Tsezic languages, and lost in others (such as Botlikh, Bagvalal, Bezhta and Hunzib). In the languages where the pattern was retained, the verbs underwent certain phonological changes before they became suffixes. In Avar, the pattern was probably lost when Avar lost the proto-Avar-Andic-Tsezic verb ‘say’ and was renewed by using a new verb ‘say’. This is why Avar has the periphrastic form and uses another verb as compared to all other languages.

The second scenario, that of a contact-induced diffusion, suggests that the ‘say’-pattern emerged in one of the Avar-Andic-Tsezic languages and spread through horizontal transmission due to the contact between these languages⁷. It seems certain that what was copied was a construction (periphrastic form) rather than a synthetic form. In order to be copied and transferred, the structural pattern should be visible and transparent. The synthetic form with the verb ‘say’ reduced to a suffix could hardly have been recognized by the speakers as a structural pattern that can be copied. Since the languages borrowed the structure rather than the matter, each utilized its own ‘say’-verb and its own imperative affix.

In the latter case, it is unclear which language could have been the source of the pattern. Socio-linguistically, the most likely candidate is Avar. At the moment, all Avar-Andic-Tsezic languages form an area with Avar as a *lingua franca*. Avar is spoken as a second language by almost everyone and is used for inter-ethnic communication. Its influence is quite strong, and there are instances of various borrowings from Avar to Andic and Tsezic languages. What prevents one from accepting this hypothesis is that it is in Avar that the construction looks young: the verb ‘say’ is not reduced to a suffix. It could have been the case that Avar had lost the ‘say’-construction and renewed it later, as suggested above. Another alternative is that, on the contrary, jussive construction in Avar developed under the influence of Andic and Tsezic languages. While linguistically this direction is the most plausible, socio-linguistically it is not, because most Andic and Tsezic languages are minorities and are usually not spoken as second languages even by small adjacent groups, let alone the Avar people.

The two scenarios could have been combined. One of the sub-branches could inherit the pattern from the proto-language. The pattern could have been further diffused to other languages due to contact. In any case, the synthetic form developed separately in each language or dialect.

4.2. *Non Avar-Andic-Tsezic languages*

The scenario of inheritance is excluded for the four non-Avar-Andic-Tsezic languages – Mehweb Dargwa, Chirag Dargwa, Archi and Lak. They belong to different branches,

⁷ Cf. Nasledskova & Netkachev (submitted) for the study of a diffusion of structural pattern with the verb ‘say’ for ordinal pronouns.

and other languages in these branches seem not to have ‘say’-jussives or optatives (except Rutul – Dobrushina, accepted for publication). Meanwhile, the contact origin of the ‘say’-forms in these languages seems quite probable, since they are all spoken within one contiguous geographic zone. The Dargwa language Mehweb and Lezgian language Archi (both minority languages) are surrounded by the speakers of Avar and Lak, both large influential languages in which Mehweb and Archi people were bilingual (Dobrushina 2013, Dobrushina 2019). Chirag Dargwa people border Lak, and many of them were bilingual in Lak (Ganenkova 2026). All these languages use their own verbs ‘say’ for jussives and optatives while sharing the structural pattern. I suppose that Lak or Avar or both could have been the sources of this shared grammaticalization. Mehweb Dargwa, Chirag Dargwa, and Archi have a “new” periphrastic form for jussives. Lak has a synthetic (thus probably older) form, which makes it a suitable candidate to be the source of the pattern.

It is important to note that the form with the ‘say’-element in Mehweb Dargwa, Chirag Dargwa, Archi, and Lak seems younger compared to the form in Andic and Tsezic languages. Most Andic/Tsezic ‘say’-forms developed an optative meaning. As I showed in Section 3.6, this meaning emerges at later stages of grammaticalization. This suggests that the Avar-Andic-Tsezic area might have been the starting point of the diffusion.

4.3. Diffusion of ‘say’-jussives and optatives: summary

I suggested that the structural pattern of ‘say’ jussives and optatives first spread among Avar-Andic-Tsezic languages due to inheritance, contact, or both, and then influenced non-Avar-Andic-Tsezic languages spoken in the same area.

As was shown above (Section 3.6), the development from reported speech constructions to inflectional forms is similar in all thirteen languages: change of argument marking of the verb ‘say’, loss of its meaning, phonological reduction, and emergence of optative usages. The noticeable similarity of the grammaticalization processes in different languages is in accordance with the findings of grammaticalization and language contact studies: “contact-induced grammaticalization proceeds in one direction from one stage to the next, where a new stage is built on the stage immediately preceding it” (Heine & Kuteva 2020: 94).

5. Conclusion

The pattern of grammaticalization of jussives from the verb ‘say’ covers a large territory of at least 13 languages from different branches of the Nakh-Daghestanian family. I have shown that what was shared by all these languages and dialects was a reported speech construction with two imperatives (*verb.IMP say.IMP*). The study demonstrated that tracing the suffix back to its lexical source is not sufficient to uncover its origin. In line with the study by Spronck & Casartelli (2021), which deals with the grammaticalization of reported speech rather than the verb ‘say’, the case of grammaticalization described in this paper distinctly showed that the reported speech construction was at the source of jussives and optatives. The traces of two predicates can be seen in the stems and the bound suffixes of inflectional jussives and optatives. Some processes of grammaticalization had already taken place at the level of the constructions, including the changes in the argument coding and semantics. As formulated by Bybee (2006: 730), “grammar is built up from specific instances of use that marry lexical items with constructions”.

In the course of grammaticalization, the verb ‘say’ underwent desemanticization. Just as quotative markers are usually attested in complement clauses with verbs of thinking and knowledge or emotional predicates, as well as clauses expressing cause and purpose, ‘say’-jussives and optatives drifted away from the reportative construction. Firstly, they obtained the ability to combine with inanimate subjects and denote indifference. Later, they developed the meaning of the speaker’s wish.

In the study of the contexts where the speech verbs are used in functions that are not explicable in terms of their primary meaning in the languages of Siberia, Matic’ & Pakendorf (2013) come to the conclusion that practically all functions of speech verbs can be reduced to one common semantic/pragmatic feature labelled *dissociation* which highlights the fact that the speaker distances him/herself from the representation of a spoken or mental text (see the definition in Güldemann 2008: 6). This observation does not hold for jussives. Although the speaker delegates the command to the third person or the fulfilment of the blessing or curse to the supernatural forces, the dissociation is not in the focus of this pattern. The verb ‘say’ in jussives plays a role which in many other languages is played by causative verbs, such as *let* in English, *pustit’* (‘let’) in Russian or the causative morpheme in Chechen. The role of the verb ‘say’ is therefore close to that of ‘order’ or ‘request’. This was noticed in the typological study of the most frequent diachronic processes of

emergence of positive directive strategies by Mauri & Sanso (2011). In their sample of 200 languages, they found only one language that used the verb ‘say’ as a source for the jussive; unsurprisingly, it was the Nakh-Daghestanian language, Chamalal. Mauri & Sanso (2011) qualified this case as a subtype of causative-based jussives.

While the jussives based on causative-proper verbs are common, ‘say’-based jussives seem rare. The factor that enhanced its wide expansion in Daghestan might be related to the very fact that the ‘say’-forms serve for the indirect command. As was repeatedly observed, pragmatically motivated linguistic items with interactional meaning are easily borrowed and quickly diffused between languages (Aikhenvald 2007: 261, Andersen 2014, Daniels & Brooks 2019: 536). The indirect commands are prominent discourse forms; frequently encountered in dialogues, they are easily spotted, copied, and used.

The combination of inheritance and contact led to the wide diffusion of the typologically infrequent grammaticalization pattern with reported imperatives used for jussives. The pattern emerged in the Avar-Andic-Tsezic proto-language or was already at the level of descendant languages, was maintained, lost, or renewed due to contact, and spread to the non-Avar-Andic-Tsezic languages, resulting in a grammaticalization area spanning a large territory of the Central and Northern Daghestan.

Acknowledgements

First of all, I would like to express my sincere gratitude to the native consultants. Without their willingness to share their knowledge, this paper would not have been possible. I am also grateful to Michael Daniel, Timur Maisak, and Konstantin Filatov for discussing some of the preliminary findings of the study with me. Finally, I thank the editors and anonymous reviewers for their helpful comments on this paper.

Abbreviation

1,2,3,4,5 = gender markers	F = feminine	LAT = motion into a spatial domain
ABS = absolutive	GEN = genitive	LCT = locutor
ALL = allative	HPL = human plural	LOC = locative
APUD = spatial domain near a hallmark	IMP = imperative	LV = light verb
ATTR = attributivizer	IMP1 = imperative in transitive imperative utterances where p is expressed by a first person pronoun	NOM = nominative
CAUS = causative	IN = spatial domain inside a (hollow) landmark	NPL = non-human plural
COMIT = comitative	INCL = inclusive	OBL = oblique
COND = conditional	INTER = spatial domain between multiple landmarks	OPT = optative
CONT = spatial domain in contact with a hallmark	IPFV = imperfective	PFV = perfective
CVS = converb	IRR = irrealis	PL = plural
DAT = dative	JUSS = jussive	PROH = prohibitive
DEM = demonstrative		PRS = present
EMPH = emphatic		PTCP = participle
ERG = ergative		PV = preverb
EXP = experiential		SG = singular

References

- Abdullaev, Isa X. 2015. *Laksko-russkij slovar'* [Lak-Russian dictionary]. Maxačkala: IJaLI DNC RAN, ALEF.
- Aikhenvald, Alexandra Y. 2007. Semantics and pragmatics of grammatical relations in the Vaupés linguistic area. In Alexandra Y. Aikhenvald & Robin M. Dixon (eds.), *Grammars in contact: A cross-linguistic typology*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780199207831.003.0010>.
- Aikhenvald, Alexandra Y. 2010. *Imperatives and Commands*. Oxford: Oxford University Press.
- Alekseev, Mixail E., Ataev, Boris M., Magomedov, Magomed A., Magomedov, Magomed I., Madieva, Gulzhagan I., Saidova, Patimat A. & Samedov, Dzhalil S. 2012. *Sovremennyy avarskij jazyk* [Contemporary Avar language]. Makhachkala: Aleph.
- Ammann, Andreas & van der Auwera, Johan. 2004. Complementizer-headed Main Clauses for Volitional Moods in the Languages of South-Eastern Europe: A Balkanism? In Olga Mišeska Tomić (ed.), *Balkan Syntax and Semantics*, 293–314. Amsterdam; Philadelphia: John Benjamins. <https://doi.org/10.1075/la.67.15com>

- Andersen, Gizle. 2014. Pragmatic borrowing. *Journal of Pragmatics* 67. 17–33. <https://doi.org/10.1016/j.pragma.2014.03.005>.
- Arkadiev, Peter & Timur Maisak. 2018. Grammaticalization in the North Caucasian languages. In Heiko Narrog & Berndt Heine (eds.), *Grammaticalization from a typological perspective*, 116–145. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780198795841.003.0007>
- Armbruster, Charles Huber. 1960. *Dongolese Nubian*. Cambridge: Cambridge University Press.
- Authier, Gilles. 2012. *Grammaire juhuri, ou judéo-tat, langue iranienne des Juifs du Caucase de l'est*. Wiesbaden: Dr. Ludwig Reichert Verlag Wiesbaden. <https://doi.org/10.29091/9783752005349>.
- Authier, Gilles. 2015. *Optatives and indirect imperatives in languages of Daghestan*. Presentation at the University of Chicago, May 14-15.
- Bernot, Denise. 1980. *Le prédicat en birman parlé*. (Langues et civilisations de l'Asie du Sud-Est et du monde insulindien, 8). Paris: SELAF.
- Bybee, Joan L., Revere Dale Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*, Vol. 196. Chicago: University of Chicago Press.
- Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733. <https://doi.org/10.1353/lan.2006.0186>.
- Chumakina, Marina. 2019. Bound but still independent: quotative and verificative in Archi. In Matthew Baerman (ed.), *Morphological perspectives: Papers in honour of Greville G. Corbett*, 281–302. Edinburgh: Edinburgh University Press. <https://doi.org/10.1515/9781474446020-014>.
- Collins, Peter. 2004. Let-imperatives in English. *International Journal of Corpus Linguistics* 9(2). 299–319. <https://doi.org/10.1075/ijcl.9.2.07col>.
- Dailidénaitė, Milda. 2022. The Livonian jussive: person and function. Eesti ja soome-ugri keeleteaduse ajakiri. *Journal of Estonian and Finno-Ugric Linguistics* 13(1). 65–90. <https://doi.org/10.12697/jeful.2022.13.1.03>.
- Daniels, Don & Joseph Brooks. 2019. The History of* = a: Contact and Reconstruction in Northeast New Guinea. *Journal of Language Contact* 12(3). 533–568. <https://doi.org/10.1163/19552629-01203001>.
- Daniel, Michael. 2015. Logophoric reference in Archi. *Journal of Pragmatics* 88. 202–219. <https://doi.org/10.1016/j.pragma.2015.07.002>.

- Daniel, Mikhail, and Timur Maisak. 2014. Grammatikalizacija verifikativa. Ob odnoj agul'sko-arčinskoj paralleli. [Grammaticalization of verificative: on one parallel between Aghul and Archi]. In Mikhail Daniel, Ekaterina Ljutikova, Vladimir Plungian, Sergej Tatevosov & Olga Fedorova (eds.), *Konstanty. Peremennyje. Pamjati A.E. Kibrika*, 377–406. St. Petersburg: Aletya.
- Dobrushina, Nina. 2011. The Optative domain in East Caucasian languages. In Gilles Authier & Timur Maisak (eds.), *Tense, Aspect, Modality and Finiteness in East Caucasian Languages*, 95–130. Bochum: Brockmeyer.
- Dobrushina, Nina. 2012. What is the Jussive for? *Linguistics* 50(1). 1–25.
- Dobrushina, Nina, Johan van der Auwera & Valentin Goussev. 2013. The Optative. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *WALS Online* (v2020.4) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.13950591> (Available online at <http://wals.info/chapter/73>, Accessed on 2026-05-08.).
- Dobrushina, Nina. 2019. The Language and people of Mehweb. In Michael Daniel, Nina Dobrushina & D. Ganenkov (eds.), *The Mehweb language: Essays on phonology, morphology and syntax*, 1–15. Language Science Press.
- Dobrushina, Nina. 2021. Optatives. In Michael Daniel, Konstantin Filatov, Timur Maisak, George Moroz, Timofey Mukhin, Chiara Naccarato & Samira Verhees (eds.), *Typological Atlas of the Languages of Daghestan* (TALD). Verhees. Moscow: Linguistic Convergence Laboratory, NRU HSE. 10.5281/zenodo.6807070. <http://lingconlab.ru/dagatlas>.
- Dobrushina, Nina. 2025. Jussive. In George Moroz, Michael Daniel, Konstantin Filatov, Timur Maisak, Timofey Mukhin, Irina Politova, Elena Shvedova, Samira Verhees and Chiara Naccarato, *Typological Atlas of the Languages of Daghestan* (TALD). Moscow: Linguistic Convergence Laboratory, HSE University. 10.5281/zenodo.6807070. <https://lingconlab.ru/tald>.
- Èldarova, Roza G. 1999. *Lakskij glagol (sistema glagol'nogo slovoizmenenija): Učebnoe posobie* [Verb in Lak (conjugation). Didactic materials]. Makhachkala: Dagestan State University.
- Epps, Patience. 2010. Linking valence change and modality: Diachronic evidence from Hup (Amazonia). *International Journal of American Linguistics* 76(3). 335–356. <https://doi.org/10.1086/652792>.
- Haspelmath, Martin. 1993. *A grammar of Lezgian*. Berlin/New York: Mouton de Gruyter. <https://doi.org/10.1515/9783110884210>.

- Heine, Bernd & Tania Kuteva. 2020. Contact and grammaticalization. In Raymond Hickey (ed.), *The Handbook of Language Contact*, 2nd ed., 93–112. Chichester: Wiley Blackwell. <https://doi.org/10.1002/9781119485094.ch4>.
- Filatov, Konstantin & Michael Daniel (eds.). 2021. *DagSwadesh: 100 Swadesh lists from Daghestan. An online database of basic vocabulary divergence across neighbour villages*. Moscow: Linguistic Convergence Laboratory, HSE University. <https://github.com/LingConLab/dagswadesh>.
- Friedman, Victor A. 2020. Lak. In Maria Polinsky (ed.), *The Oxford handbook of languages of the Caucasus*, 201–241. Oxford University Press, USA. <https://doi.org/10.1093/oxfordhb/9780190690694.013.11>.
- Ganenkov, Dmitry. 2026. *A grammar of Chirag Dargwa*. Unpublished manuscript.
- Grosz, Patrick G. 2012. *On the grammar of optative constructions*. Amsterdam: John Benjamins Publishing Company. <https://doi.org/10.1075/la.193>.
- Güldemann, Tom. 2008. *Quotative Indexes in African Languages: A Synchronic and Diachronic Survey*. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110211450>.
- Gusev, Valentin. 2013. *Tipologija imperativa* [A typology of imperative]. Moskva: Jazyki slav'anskoj kul'tury.
- Khalilova, Zaira. 2009. *A grammar of Khwarshi*. LOT, University of Utrecht, The Netherlands.
- Kibrik, Aleksandr E. 1977. *Opyt strukturnogo opisaniya arčinskogo jazyka. Tom II. Taksonomičeskaja grammatika* [Structural description of Archi, Vol. 2. Taxonomic grammar]. Moscow: Izdatel'stvo Moskovskogo universiteta.
- Kibrik, Aleksandr E. & Sandro V. Kodzasov. 1990. *Sopostavitel'noe izučenie dagestanskix jazykov. Imja. Fonetika* [A comparative study of Daghestanian languages. Noun. Phonetics]. Moscow: Izdatel'stvo Akademii nauk SSSR.
- Koryakov, Yuri. 2006. *Atlas kavkazskix jazykov* [Atlas of the languages of the Caucasus]. Moscow: IJa RAN.
- Kuteva, Tania, Heine, Bernd, Hong, Bo, Long, Haiping, Narrog, Heike, & Rhee, Seongha. 2019. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781316479704>.
- Lehmann, Christian. 2020. Univerbation. *Folia Linguistica* 54 (s41-s1). 205–252. <https://doi.org/10.1515/flih-2020-0007>.

- Magomedova, Patimat T., 2004. Chamalal. In Job, Michael (ed.), *The Indigenous Languages of the Caucasus, The North East Caucasian Languages*, Part 1, Vol. 3, 1–65. Caravan Books, Ann Arbor.
- Magomedova, Patimat T. 1999. *Čamalinsko-russkij slovar'* [Chamalal-Russian dictionary]. Makhachkala: IJaLI.
- Maisak, Timur A. 2016. Morphological fusion without syntactic fusion: The case of the “verificative” in Agul. *Linguistics* 54(4). 815–870. <https://doi.org/10.1515/ling-2016-0015>.
- Matić, Dejan & Pakendorf, Brigitte. 2013. Non-canonical SAY in Siberia: Areal and genealogical patterns. *Studies in Language* 37(2). 356–412. <https://doi.org/10.1075/sl.37.2.04mat>.
- Mauri, Caterina & Andrea Sansò. 2011. How directive constructions emerge: Grammaticalization, constructionalization, cooptation. *Journal of Pragmatics* 43(14). 3489–3521. <https://doi.org/10.1016/j.pragma.2011.08.001>.
- Mudrak, Oleg A. 2019. *Avarskie osnovy* [Avar stems]. Moskva: Jazyki narodov mira.
- Mudrak, Oleg A. 2020. *Andijskie osnovy* [Andic stems]. Moskva: Jazyki narodov mira.
- Nasledskova, Polina & Ivan Netkachev. 2026. The ‘say’ strategy for ordinal numerals in East Caucasian: conflicting evidence on the evolution of a rare morphosyntactic pattern. *Studies in Language. International Journal Sponsored by the Foundation “Foundations of Language”*. Online first article. <https://doi.org/10.1075/sl.24077.nas>.
- Nichols, Johanna. 2003. The Nakh-Daghestanian consonant correspondences. In Dee Ann Holisky and Kevin Tuite (eds.), *Current trends in Caucasian, East European and Inner Asian linguistics: papers in honor of Howard I*, 207–264. Aronson. Amsterdam: John Benjamins. <https://doi.org/10.1075/cilt.246.14nic>.
- Nikolaev, Sergei & Sergei Starostin. 1994. *A North Caucasian Etymological Dictionary*. Moscow: Asterisk.
- Nikolaeva, Irina. 2016. Analyses of the semantics of mood. In J. Nuyts, & J. van der Auwera (eds.), *The Oxford Handbook of Modality and Mood*, 68–85. Oxford University Press.
- Ochs, Samira. 2021. Prohibitive and its relation with the imperative. In Michael Daniel, Konstantin Filatov, Timur Maisak, George Moroz, Timofej Mukhin, Chiara Naccarato & Samira Verhees (eds.), *Typological Atlas of the Languages of Daghestan* (TALD). Moscow: Linguistic Convergence Laboratory, NRU HSE. DOI: 10.5281/zenodo.6807070. <http://lingconlab.ru/dagatlas>.

- Panova, Anastasia B. 2018. Dve klauzy v odnom slove: predvaritel'naja tipologija morfoložičeski svjazannoj komplementacii [Two clauses in a single word: preliminary typology of morphologically bound complementation]. *Typology of Morphosyntactic Parameters* 1(2). 84–99.
- Rusakov, Aleksandr & Maria Morozova. 2022. Optative in Albanian. *Syntaxe & Sémantique* 22(1). 29–55. <https://doi.org/10.3917/ss.022.0029>.
- Spronck, Stef & Daniela Casartelli. 2021. In a manner of speaking: how reported speech may have shaped grammar. *Frontiers in Communication* 6. 1–22. <https://doi.org/10.3389/fcomm.2021.624486>.
- Sumbatova, Nina R. 2019. Ličnyj dejksis v citacionnyx konstrukcijax darginskogo jazyka. (Personal deixis in the reported speech: the case of Tanti Dargwa). *Tomskij žurnal lingvističeskix i antropologičeskix issledovanij* 3(25), 72–81. <https://doi.org/10.23951/2307-6119-2019-3-25-72-84>.
- Xrakovskij, Victor S. (ed.) 2001. *Typology of Imperative Constructions*. München: Lincom Europa. (Revised and expanded translation of Xrakovskij, Viktor S., ed. 1992. *Tipologija imperativnyx konstrukcij*. St. Petersburg: Nauka.).
- Zerzele, Vasilij. 2024. Plural marking in imperatives and prohibitives. In Michael Daniel, Konstantin Filatov, Timur Maisak, George Moroz, Timofej Mukhin, Chiara Naccarato & Samira Verhees (eds.), *Typological Atlas of the Languages of Daghestan* (TALD). Moscow: Linguistic Convergence Laboratory, NRU HSE.10.5281/zenodo.6807070. <http://lingconlab.ru/dagatlas>.

CONTACT

nina.dobrushina@gmail.com

Basic verbs in European languages

ÅKE VIBERG

UPPSALA UNIVERSITY

Submitted: 24/08/2025 Revised version: 05/02/2026

Accepted: 13/04/2026 Published: 24/07/2026



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

Abstract

This article presents a corpus-based approach toward a lexical typology of verbs in European languages, based on a comparison of the meanings of the most frequent verbs classified into semantic fields. The comparison is made in two steps. Among the twenty most frequent verbs, there is a set of grammatical verbs that top the list (with a few exceptions): ‘be’, ‘have’ and modals meaning ‘can’ and ‘must’. Among the lexical verbs, there are nine nuclear verbs with meanings that are distributed between the most basic verbal semantic fields (indicated within parentheses): ‘go’, ‘come’ (Motion); ‘give’, ‘take’ (Possession), ‘make’ (Production), ‘say’ (Verbal communication), ‘see’ (Perception), ‘know’ (Cognition), ‘want’ (Desire). Comparisons are made in two ways based on the rank order of the verbs according to descending frequency and according to a frequency measure, based on a logarithmic scale. As a second step, the comparison is extended to the 100 most frequent verbs. The result is compared to the inventories of simple verbs in some non-European languages with a very restricted number of such verbs. The frequency-based comparisons are complemented with the result of corpus-based contrastive studies of selected aspects of the semantic structure of the semantic fields to which the nuclear verbs belong.

Keywords: lexical typology; semantic fields; verbs; basic vocabulary; word frequencies; European languages; corpus-based contrastive study

1. Introduction

The general aim of this article is to find a model for comparing the basic structure of whole vocabularies in order to provide a framework for contrastive and typological

investigations of the semantic organization of lexical structure from a holistic perspective. The study focuses on verbs. In principle, the approach is corpus-based combining frequency counts of large corpora with corpus-based (or usage-based) studies of the basic words that have been identified and the semantic fields to which they belong. The use of frequency is motivated by the central role it plays for language processing and learning (Divjak 2019).

European languages have an open class of verbs which tends to contain something in the range of 10,000 well-established verbs (lexemes). Despite the vast number of words that are potentially available for describing events of various types, it turns out that a small number of verbs are used on most occasions. In Swedish, for instance, the twenty most frequent verbs cover 49% of all the occurrences of verbs in running text, the 50 most frequent verbs cover 60% and the 100 most frequent verbs 68% (Viberg 2006a: 105). Furthermore, there is a striking similarity across languages with respect to the meaning of these verbs. This was shown in Viberg (1993) by comparing the basic meanings of the 20 most frequent verbs in 11 European languages. There are nine lexical meanings that tend to be realized as one of the twenty most frequent lexical verbs in all European languages. These meanings are distributed between the most basic verbal semantic fields (indicated within parentheses): ‘go’, ‘come’ (Motion); ‘give’, ‘take’ (Possession), ‘make’ (Production), ‘say’ (Verbal communication), ‘see’ (Perception), ‘know’ (Cognition), ‘want’ (Desire). For the sake of simplicity, these verb meanings (and only these) will be referred to as “nuclear verbs”. Even when a verb with this meaning does not appear among the 20 most frequent verbs, it is still counted as a nuclear verb but as an exception. Some non-European languages lack direct equivalents of some of the nuclear verbs and the closest correspondences are analytical constructions (see, for example, Section 7). The appearance of ‘be’ and ‘have’ and modal verbs among the most frequent verbs is an areal-specific characteristic of European languages.

The nuclear verbs tend to have many hyponyms, to have characteristic and elaborate patterns of polysemy and to have a strong tendency to develop grammatical meanings. As a set, they cover most of the different types of argument structures characteristic of an individual language and in several respects serve as syntactic prototypes. They also play a central role in first and second language acquisition and in lexical regression such as attrition and certain types of language loss (Viberg 1993, 2002).

Frequency will be used to identify the most basic verbs in an individual language. In a second step, I try to answer the question to what extent basic verb meanings are

shared. Basic verbs that are shared with only a few languages are identified as language-specific. It would be possible to simply talk about the most frequent verbs, but it is convenient to refer somewhat more loosely to basic verbs. Basicness is not an absolute concept. A large part of this paper will deal with the 20 most frequent verbs, but reference will also be made to the 100 most frequent verbs, which are basic but to a lower degree. The cut-off point is in principle arbitrary. What is essential is to try to find the most informative cut-off point to answer a specific research question and then to stick to this cut-off point throughout the discussion. What is referred to as basic verbs has sometimes been referred to as core verbs or nuclear verbs in the literature. To avoid confusion, it is important to underline that the term “nuclear verb” in this paper will be used exclusively to refer to the specific set of nine verbs, which are basic in European languages and tend to belong to the 20 most frequent verbs.

2. Testing the generalization

At the time when the earlier study (Viberg 1993) was carried out, most frequency counts were based on corpora containing around one million words. As will be shown below, the general conclusions turn out to hold for the much larger corpora available today such as the British National Corpus (BNC) containing around 100 million words taken from a wide range of spoken and written genres. In the following, the conclusions from the earlier study will be tested against data from corpora that are available today. What is presented should be regarded as an exploratory study, since there is a need for further development of the methodology and more detailed analysis of each of the verbs that have been identified as basic. To my knowledge, however, there is no contrastive-typological study of basic vocabulary following the same kind of approach.

2.1. Language sample and data sources

The present study is restricted to European languages. The major source of data consists of the frequency lists or dictionaries in Table 1. (To simplify reference, all sources will be referred to as “lists”.) The lists are based on several million words from reasonably balanced corpora with a few exceptions. In principle, there are at least two examples of each of the Germanic, Romance, Slavic and Uralic languages. In addition, there is Greek and Welsh. Lithuanian is included as an example of a Baltic language

despite the restricted size of the corpus (1 million words) and the dominance of formal registers (scientific 32%, fiction 12%, official documents 10%, periodicals 46%). The only major groups that are missing are Albanian and Basque. Swedish and Finnish are included because data from my earlier contrastive studies based on these languages will be used to discuss problematic semantic issues, even though the Swedish corpus contains only one million words and the Finnish corpus only is based on News. Code in Table 1 refers to the abbreviated forms of the names of the languages that will be used in several of the tables to save space. Information is also given about the classification in Glottolog (Hammarström et al. 2026). In the column “Language”, the glottocode appears under the conventional English name of the language (see <https://glottolog.org/glottolog/language>; accessed 2026-07-06).

Code	Language	Size Million words	Composition	Reference
INDO-EUROPEAN				
<i>Germanic</i>				
ENG	English stan1293	100	Balanced Spoken & Written	BNC (Kilgarriff 1998)
GER	German stan1295	20	Balanced Spoken & Written	Jones & Tschirner (2020) 2nd edn
SWE	Swedish swed1254	1	Balanced Written	The Stockholm-Umeå Corpus (SUC 1997)
<i>Romance</i>				
FRE	French stan1290	23	Balanced Spoken & Written	Lonsdale & Le Bras (2009)
SPA	Spanish stan1288	20	Spoken, Fiction, Non- fiction	Davies (2006)
<i>Greek</i>				
GRE	Greek mode1248	47	Written language	Hellenic National Corpus (data obtained from Maria Gavrilidou December 2009 from the version that existed at that time)
<i>Slavic</i>				
CZE	Czech czec1258	20	Spoken, Fiction, Non- fiction, Newspaper	Čermák & Křen (2011)
RUS	Russian russ1263	150	Internet corpus	Sharoff et al. (2013)

(continued)

Code	Language	Size Million words	Composition	Reference
INDO-EUROPEAN				
Baltic				
LIT	Lithuanian lith1251	1	Written language (only 12% fiction)	Dažninis rašytinės lietuvių kalbos žodynas (Andrius Utka 2009)
Celtic				
WEL	Welsh wels1247	11	Spoken and written	Knight et al. (2020) https://doi.org/10.17035/d
URALIC				
Finnic				
EST	Estonian esto1258	15	Fiction, Press, Science (5 mill words each)	The Frequency List of Estonian Literary Language
FIN	Finnish finn1318	44	Press ("printed news")	Wiktionary:Frequency_lists/Finnish https://en.wiktionary.org/wiki/ (Accessed 2026-07-06)
Hungaric				
HUN	Hungarian hung1274	777	Fiction, Non-fiction, Spoken	Hungarian Gigaword Corpus (Oravecz et al. 2014)

Table 1. Language sample and data sources.

2.2. Methodological problems

The frequency lists are developed according to different criteria in many respects, but with few exceptions data are accepted as they are given in the sources without any attempt to compensate for such discrepancies.

2.2.1. Variation across genres/registers

Although the frequency lists are based on “balanced” corpora, it is problematic to decide which genres should be distinguished and what the proportions should be between various sub-corpora. It should also be remarked that most of the discussion in this paper will be restricted to the standard languages. Another restriction is that spoken language is not included in all corpora. This is particularly problematic when the result is compared with typological studies, which typically rely on spoken data. To some extent this can be compensated for by using data from subtitle corpora, which

are available for all European languages, although this has not been accounted for in the present study. This issue is discussed in Viberg (2020).

2.2.2. *Raw frequency and dispersion*

The frequency of a word varies in different types of text. How widely a word is spread across the documents in a corpus can be measured with various measures of dispersion. Even distribution of a word gives a high score, whereas a low score indicates that a word appears only within a restricted range of texts. The French list (Lonsdale & Le Bras 2009) shows the raw frequency of each word, whereas the rank order is based on a combination of the raw frequency and a measure of dispersion called deviation of proportions (DP) (Gries 2008). Ranking that takes dispersion into consideration is more accurate but is available only for some languages.

2.2.3. *Lemmatization*

The lexical items consist of single words but derived and compound words are sometimes treated differently in the lists. Morphological complexity forms a continuum: Simple root > Derived word > Compound word > Lexicalized phrase. Kopecka (2006) shows how derivation forms a continuum with respect to French motion verbs (e.g., *dé-rouler* ‘roll out’ vs. *sortir* ‘go = OUT’). On compound verbs, see Viberg (2017a) and on lexicalized phrases or idioms as a continuum, see Langlotz (2006).

In general, the identification of lemmas in the frequency lists has been followed. In some cases, the frequency lists treat derived forms in a way that reduces comparability. A clear case is the negation in Lithuanian which is represented as a prefix of the verb. Such forms are treated as lemmas in the frequency list, but in this paper the frequencies of the positive and the negative forms have been collapsed and are represented by the positive form of the verb. This appears to be unproblematic in most cases. In Russian and Czech, the frequencies of the perfective and imperfective forms of an aspectual verb pair are indicated separately but have been collapsed. This is not unproblematic but appears to improve comparability. Rank order on the other hand is indicated for the most frequent member of the pair.

2.2.4. Glossing

The identification of the meanings has been based primarily on the glosses given in the frequency dictionaries. For some languages, the source has been a frequency list that does not provide any glosses, and in that case the meaning has been identified with a bilingual dictionary (X to English). When several glosses have been provided, one meaning has been identified as the basic meaning. A further resource that has been used is the Word Alignment Database (WAD) included as a tool in OPUS (Tiedeman 2012). This tool takes a word form in one language as input and shows the word forms that appear in the translations according to a formula. This tool has been used frequently in this study, but it would take up too much space to account for all these searches. See Appendix 1 for an example. The use of translation to identify correspondences across languages in contrastive studies is discussed in Viberg (to appear).

2.3. Basing the comparison on frequency rank order

Table 2 shows the rank of the nine nuclear verbs together with BE, HAVE, CAN and MUST. Ranks higher than 20, which represent exceptions, are shaded.

	BE	HAVE	CAN	MUST	MAKE	GO	COME	GIVE	TAKE	SAY	SEE	KNOW	WANT
ENG	1	2	5	23	9	7	16	17	12	6	10	11	21
GER	1	2	4	6	9	12	11	7	26	5	14	13	8
SWE	1	2	3	16	9	11	6	15	13	8	10	19	14
FRE	1	2	3	7	4	10	18	9	8	6	14	12	11
SPA	1(3)	4(2)	6	15	5	8	26	10	21/30	7	9	11	12
GRE	1	2	4	8	7	18	27	9	13	3	12	34	11
CZE	1	2	3	5	14	6	15	10	19	9	12	4	8
RUS	1	---	2	ADJ	13	11	19	10	33	3	15	6	7
LIT	1	3	2	4	22	36	57	64	18	9	28	7	6
WEL	1	---	9	N	3	4	5	10	25	8	7	13	37
EST	1	---	4	(3)	6	8	5	12	11	7	14	9	15
FIN	1	---	4	?	6	20	5	12	11	8	27	30	13
HUN	1	---	(4)	2	6	22	23	10	7	5	9	4	18

Table 2. The ranking of shared basic verbs in 13 European languages.

Except for Lithuanian, at most two nuclear verbs have a higher rank than 20 and the deviation is not very high in any case. Another exception is that *MUST* is not realized as a verb but as a noun in Welsh and as an adjective in Russian. These peculiarities will be discussed below in Sections 3-9.

2.4. Basing the comparison on levels of the natural logarithm (LN)

An alternative to rank order as a basis for comparison is to base the comparison directly on the frequencies. First the raw frequencies are normalized and specified as frequencies per million words (pm). However, the problem remains to find natural cut-off points. One solution is to use a logarithmic scale and to that end the frequencies of the basic verbs are expressed as values of the natural logarithm (with the Latin acronym LN) in Table 3. (The base of LN is approximately equal to 2.718.) At the bottom of the table the result is summed up by an indication of the median and the highest and the lowest value for each verb.

	BE	HAVE	CAN	MUST	MAKE	GO	COME	GIVE	TAKE	SAY	SEE	KNOW	WANT
ENG	10.7	9.5	8.4	6.6	7.7	7.8	7.3	7.2	7.5	8.1	7.6	7.5	6.9
GER	10.1	9.4	8.6	7.9	7.6	7.4	7.5	7.7	6.4	8.0	7.2	7.2	7.6
SWE	10.1	9.5	8.9	7.1	7.9	7.8	8.1	7.2	7.7	8.0	7.9	6.9	7.3
FRE	10.0	9.8	8.1	7.6	8.4	7.7	6.8	6.8	7.0	8.0	7.2	7.3	7.4
SPA	9.8	8.6	8.3	7.0	8.3	8.2	6.7	7.6	6.5	8.3	7.3	7.5	7.3
GRE	9.4	9.1	7.9	7.6	7.7	6.6	6.4	7.4	6.9	8.1	6.9	6.3	7.1
CZE	10.3	8.6	7.7	7.1	6.9	7.1	6.3	7.0	6.2	7.7	6.6	7.4	7.0
RUS	9.4	---	8.2	ADJ	7.4	7.2	6.5	6.9	6.5	8.2	7.1	7.4	7.1
LIT	9.9	8.4	8.7	7.6	7.3	6.9	6.0	5.9	6.9	7.7	7.0	7.4	7.4
WEL	11.0	---	7.7	N	8.5	8.0	8.0	7.6	6.5	7.7	7.8	6.9	6.2
EST	10.7	---	8.3	(8.3)	8.0	7.8	8.1	7.5	7.5	7.8	7.3	7.6	7.3
FIN	11.0	---	8.4	?	8.3	7.0	8.4	7.4	7.4	8.0	6.9	6.8	7.3
HUN	9.3	---	(7.9)	8.1	7.3	6.6	6.5	6.9	7.1	7.7	7.0	7.9	6.8
<i>Med</i>	10.1	---	8.3	(7.1)	7.7	7.4	6.8	7.2	6.9	8.0	7.2	7.4	7.3
<i>High</i>	11.0	9.8	8.9	8.1	8.5	8.2	8.4	7.7	7.7	8.3	7.9	7.9	7.6
<i>Low</i>	9.3	8.4	7.7	6.6	6.9	6.6	6.0	5.9	6.2	7.7	6.6	6.3	6.2

Table 3. The LNs of shared basic verbs in 13 European languages.

BE has an LN around 10 (Median: 10.1. Lowest: 9.3 – Highest 11.0) followed by HAVE around 9 in the languages where the verb is established (8.4 – 9.8). CAN reaches above 8.0 in most cases (7.7 – 8.9), but MUST is more variable and has a somewhat lower value and is not realized as a verb in some languages. The rest of the verbs in Table 3 make up the nuclear verbs that tend to have an LN above 7 but below 8.0, which corresponds to a frequency between 1,100 and 3,000 per million words (pm). Thus, the use of LN shows how the most frequent verbs tend to appear within specific levels.

2.5. A brief look at the 100 most frequent verbs

The discussion will account systematically for all the nuclear verbs that tend to appear among the 20 most frequent verbs. In addition, several verbs with a rank between 21 and 100 will be commented on below when they belong to the same field as one of the nuclear verbs. Table 4 compares the LNs at various frequency ranks. The median LN for the verb that has rank 100 is 5.4, which is still relatively high and corresponds to a frequency close to 220 pm.

	BE	HAVE	Rank 5	Rank 10	Rank 20	Rank 50	Rank 100
English	10.7	9.5	8.4	7.6	6.9	5.9	5.4
German	10.1	9.5	8.0	7.6	6.6	6.0	5.2
Swedish	10.1	9.5	8.4	7.9	6.9	6.1	5.4
French	10.0	9.8	6.8	7.7	6.3	5.4	4.7
Spanish	9.8	9.2	8.3	7.7	7.0	5.8	5.5
Greek	9.4	9.1	7.8	7.2	6.6	6.0	5.5
Czech	10.3	8.6	7.1	7.0	5.7	5.0	4.5
Russian	9.4	---	8.2	6.9	6.5	5.9	5.4
Lithuanian	9.9	8.4	7.6	6.9	6.7	6.0	5.6
Welsh	11.0	---	8.0	7.6	6.5	6.1	5.4
Estonian	10.7	---	8.1	7.6	7.0	6.1	5.6
Finnish	11.0	---	8.4	7.4	7.1	6.5	6.0
Hungarian	9.3	---	7.7	6.9	6.6	5.8	5.2
<i>Median</i>	10.1	(9.4)	8.1	7.6	6.6	6.0	5.4
<i>Highest</i>	11.0	9.8	8.4	7.9	7.1	6.5	6.0
<i>Lowest</i>	9.3	8.4	7.1	6.9	5.7	5.0	4.5

Table 4. A comparison of the LNs at various frequency ranks.

The disposition of the rest of the paper is as follows. Section 3-9 will be devoted to a discussion of the set of shared basic verbs of European languages, organised in terms of semantic fields. This is followed by a couple of more general sections. The semantics of the nuclear verbs will be primarily discussed based on my own corpus-based contrastive studies. Other work will also be referred to but often rather briefly for reasons of space. More references are given in my earlier studies, which are based primarily on the English-Swedish Parallel corpus (Altenberg & Aijmer 2000) and on the Multilingual Parallel Corpus (MPC), which consists of extracts from 22 Swedish novels and their translations into English, German, French, Spanish and Finnish. Data from further languages will be taken from OPUS, in particular Subtitles and to a lesser extent from Europarl. Several methodological problems are associated with the use of translations. For my view on this subject, see Viberg (2016a).

3. BE and HAVE

Table 5 shows the frequency (pm) and the LN of the verbs meaning BE and HAVE in the sample of European languages.

Language	BE			HAVE		
	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>
English	<i>be</i>	10.65	42396	<i>have</i>	9.53	13756
German	<i>sein</i>	10.08	23949	<i>haben</i>	9.43	12452
Swedish	<i>vara</i>	10.09	24094	<i>ha</i>	9.53	13826
French	<i>être</i>	10.02	22372	<i>avoir</i>	9.78	17610
Spanish	<i>ser</i>	9.84	18710	<i>tener</i>	8.58	5332
	<i>estar</i>	8.58	5321	<i>haber</i>	9.20	9848
Greek	<i>ίμε</i>	9.44	12539	<i>έxo</i>	9.08	8782
Czech	<i>být</i>	10.27	28850	<i>mit</i>	8.57	5271
Russian	<i>byt'</i>	9.36	11588	<i>(imet')</i>	6.96	1056
Lithuanian	<i>(ne)būti</i>	9.86	19187	<i>turėti</i>	8.41	4502
Welsh	<i>bod</i>	11.10	66137	WITH-possessive	---	---
Estonian	<i>olema</i>	10.66	42653	LOCATIVE	---	---
Finnish	<i>olla</i>	11.03	61736	LOCATIVE	---	---
Hungarian	<i>van</i>	9.33	11308	DATIVE	---	---

Table 5. BE and HAVE.¹

¹ In Lithuanian, the frequency of the positive form has been collapsed in Table 5 and the following tables with the frequency of the negative form of the verb, which is indicated with the prefix ne- and

BE-verbs have been studied extensively from a crosslinguistic perspective, at least since the six volumes edited by Verhaar (1967-1972). The seminal typological work is Stassen (1997), which is based on data from 210 languages, followed recently by Bertinetto et al. (2026a,b). In a similar way, HAVE, which is not realized as a simple verb in many languages, represents a type of predicative possession described in Stassen (2009), see also Heine (1997: 45–142), and, for Existentials, Sarda & Lena (2023) and Viberg (2010: 129–133)². The following discussion will be restricted to European languages.

3.1. BE- and HAVE-languages

BE has rank 1 in all European languages. In Spanish, there are two verbs, *ser* and *estar*, but *ser* can be regarded as the major BE-verb. As can be seen, the most important dividing line between European languages turns out to be whether there is a verb meaning HAVE or not. This motivates a special term HAVE-language for the languages that have such a verb (Isačenko 1974).

The European languages that are not HAVE-languages appear in the Eastern and Western parts of Europe and consist of Uralic languages such as Finnish, Estonian and Hungarian and of Celtic languages such as Welsh, Irish and Breton. The Baltic and Slavic languages are HAVE-languages, except Russian and Latvian, whereas Ukrainian and Belorussian are in a stage of transition according to Isačenko (1974). Somewhat unexpectedly, a verb glossed ‘have’ (*imet’*) is found among the twenty most frequent Russian verbs, but this verb is primarily used together with abstract nouns to express Abstract possession in phrases modelled on Western European languages: *imet’ talant* ‘have the talent’, *imet’ mesto* ‘take place’ (compare French *avoir lieu*).

According to Stassen (2013a), “a core area for the Have-Possessive is western and central Europe; however, we also find quite a few instances of this option in Africa and in the Americas”.

represented as a separate word in the frequency list. In Welsh, the frequency given for the infinitive *bod* has been collapsed with that of the 3SG form *mae* (2596), which is indicated separately in the frequency list, since only this form is used in the Existential and Possessive functions.

² For reasons of space, I will only be able to give reference to relevant works without being able to discuss them properly, which is unsatisfactory. I have also included rather many references to my own work where I have motivated my own analysis, since I’m not able to do that in this paper. I hope to be able to return to the topics of some of the individual sections in separate papers.

3.2. Contrastive patterns of *BE* and *HAVE*

The range of functions covered by *BE* and *HAVE* constitutes an important component of a contrastive/typological description of basic verbs in European languages. To give a simple overview of the basic functions in European languages, English examples are used in Table 6. (The range of grammatical functions varies across languages.)

Basic functions	Examples
Identity	The tall fellow over there <i>is</i> my brother.
Description:	
Nominal description	Peter <i>is</i> a teacher.
Adjectival description	Peter <i>is</i> tall.
Location	Peter <i>is</i> in the kitchen.
Existence	There <i>is</i> a fridge in the kitchen.
Possession	Peter <i>has</i> a car.
Grammatical uses	Examples
Cleft	It <i>is</i> Peter who <i>is</i> in the kitchen.
Progressive	Peter <i>is</i> snoring.
Passive	Peter <i>was</i> killed.
Perfect	Peter <i>has</i> arrived.
Modal: Obligation	Peter <i>has</i> to go now.

Table 6. Major functions of *be* and *have* in English.

The Identity construction is used to state the identity of two referents, which may have been treated as two independent referents in earlier discourse. In this construction, the NP in the complement is always referential, whereas it is non-referential in the Nominal description. Stassen (1997) refers to Adjectival description as Property predicate and Nominal description as Class-inclusion predicate.

In European languages, a predicate nominal in general requires a copula and the copula is always a verb. This situation is not unique to Europe, but outside Europe copulas can be realized as another type of word, most frequently a personal or demonstrative pronoun. Outside Europe, there are also languages where a Zero copula is the norm. A concentration of such languages appears in Australia and New Guinea according to Stassen (2013b), who also found across a world-wide sample that a Zero

copula was possible at least in some context in 175 languages but impossible in 211 languages.

The Grammatical uses listed in Table 6 occur in various combinations across European languages. For example, in a quantitative review, Čermáková et al. (2023) found a large difference with respect to the proportion of auxiliary uses of the BE verb in Czech (47%), English (26%) and Norwegian (6%). The study was based on fiction. The large proportion in Czech primarily depends on the use of the copula to form the past tense. The use as a progressive auxiliary is prominent in English but is missing in the other two languages.

The functions of BE and HAVE form a continuum that will be referred to as the Stative scale. Table 7 shows the basic contrasts in a selection of European languages exemplifying various possibilities.

	English	Swedish	French	Spanish	Greek	Russian		Finnish
						PRS	-PRS	
Identity	be	vara	être	ser	íme	Ø	byt'	olla
Description								
Stable								
Unstable								
Location				Postural				
Existence	finnas	haber						
Possession	have	ha	avoir	tener	éxo			
Abstract							imet'	

Table 7. The stative scale. Basic contrasts in the use of BE and HAVE in a selection of European languages.

In English, *be* covers all functions except Possession. Swedish is an example of a HAVE-language where BE has a restricted use in constructions that express Location and Existence. French and Greek are examples of languages where HAVE has extended into Existence. In French, *être* ‘be’ can be used to refer to Location, but a wide range of other options are frequent (Viberg 2013a: 153). In Spanish, *haber* has the second rank but is used as an auxiliary and in Existential constructions. Another verb *tener* with rank 4 functions as a verb of possession. Spanish also has a special locative copula

estar that also can be used as a descriptive copula but primarily to refer to unstable properties. (This is a complex and much debated issue, see Camacho 2012, Leonetti et al. 2015). Russian is the only European language that makes extended use of a Zero-copula in the present tense. It is also a BE-language with an emergent HAVE (*imet'*) restricted to Abstract possession. Finnish is a consistent BE-language, where the copula is used in all functions.

BE plays an exceptionally prominent role in the clausal structure of Welsh (see King 2016). In this language, verbs in general mark tense periphrastically with *bod* 'be' except for the preterite (past perfective). Thus, the construction in (1) corresponds both to the simple and the progressive present form in English.

(1) Welsh (King 2016: 169)

<i>mae</i>	<i>hi</i>	<i>'n</i>	<i>prynu</i>
be.PRS.3SG	she	CM	buy[VN]
'She buys. / She is buying.'			

Since Welsh is a VSO-language, the copula appears initially. A Complement Marker (CM) *yn* (*'n* after a vowel) is required in front of the complement. The verb in periphrastic constructions formally has the properties of a noun and is referred to as Verbal Noun (VN). The copula also has special forms to indicate whether a clause is affirmative, interrogative or negative. This means that the copula carries very much grammatical information that appears in the initial position.

There is a trade-off between the use of the copula and the realization of concepts as a certain word class. Adjectival description, where an adjective appears in predicative position, requires the use of a copula in European languages. Outside Europe, there are languages where property concepts corresponding to adjectives in European languages are primarily realized as verbs. Stassen (2013c) refers to these as verbal property concepts and to adjectives as non-verbal property concepts. Europe, in which non-verbal property concepts dominate, belongs to a mega-area comprising also Central Asia and Siberia, India the Middle East and northern Africa, whereas verbal encoding predominates in several areas such as Africa south of the Sahara, East and Southeast Africa and North America. There are also languages such as Hausa (haus1257; Afro-Asiatic, Chadic), where property-concepts are primarily realized as nouns and appear in a possessive construction. In European language, the part-of-speech of property concepts can vary, for example, the favored realization of certain

types of Physiological Conditions or Emotions: *I'm hungry, I'm thirsty, I'm afraid*; French: *j'ai faim, j'ai soif, j'ai peur*; Greek: *pináo, ðipsáo, fováme*; German: *Ich bin hungrig / Ich habe Hunger, Ich bin durstig / Ich habe Durst, Ich habe Angst*. The large differences between European languages with respect to the frequency of the copula depend on differences in the lexicalization of concepts as a certain word class, which remains to be studied in detail, and on the variable grammatical uses of the copula, which is well described in the grammars of individual languages but needs to be systematically quantified to make contrastive comparisons.

4. Modals: CAN and MUST

In the introduction to the comprehensive reference work “Modals in the languages of Europe”, Hansen & de Haan (2009: 3) present the following working definition of modals: “word-like elements which are polyfunctional in the sense that they express no less than two types of modality”. The expression “word-like” element is motivated by the fact that modals appear on a continuum from fully lexical items to inflections. The part-of-speech also varies, but for the purposes of the present paper words that are regarded as verbs in the frequency dictionaries are regarded as modal verbs³ if they fulfill the semantic criteria, even if they may be defective in some respect.

Three types of modality are specifically mentioned by Hansen & de Haan (2009): Dynamic modality basically refers to the ability (*I can read*) or internal needs (*I must eat*) of the subject to do something (internal dynamic modality) or to possibilities or necessities due to the external circumstances (external dynamic modality). Compare: *I can (know how to) play tennis* (internal), *but I cannot play today* (external), *since the tennis court is closed*. Deontic modality typically involves permission or obligation of the subject to do something (*You may / must leave now*) or more generally indicates the degree of moral desirability of the situation referred to, whereas Epistemic modality refers to the estimation (typically by the Speaker) of the degree to which the described states of affairs apply in the world along the scale: certainty – probability –

³ In English, the modals have been included even though they are not counted as verbs in the source (Kilgarriff 1998). The reason for that is that the closest correspondences in the other European languages are counted as verbs. The frequency is based on the combined frequency of the different forms, which are listed as separate lexemes in the source: *will/would, can/could, may/might* and *shall/should*.

possibility (*He must be tired / I may be wrong*). Within each of these types, a contrast can be made between Possibility and Necessity, which forms the basis of the influential typology presented in van der Auwera & Plungian (1998). For a discussion of various views on modality, see Nuyts (2014).

The frequency data used for the present study make it possible in most European languages to identify one verb centered around Possibility (symbolized CAN) and one verb centered around Necessity (symbolized MUST) that is the major realization of the concept. Table 8 shows the frequency per million words and the corresponding value of the natural logarithm for the two types of modals in 13 European languages, which makes it possible to compare the frequency ranges of the major modals with those of the nuclear verbs.

Language	CAN			MUST		
	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>
English	<i>can</i> (including <i>could</i>)	8.38	4345	<i>must</i>	6.58	721
German	<i>können</i>	8.58	5304	<i>müssen</i>	7.93	2787
Swedish	<i>kunna</i>	8.89	7265	<i>måste</i>	7.13	1251
French	<i>pouvoir</i>	8.13	3395	<i>devoir</i>	7.61	2021
Spanish	<i>poder</i>	8.25	3837	<i>deber</i>	7.01	1112
Greek	<i>boró</i>	7.94	2807	<i>prépi</i>	7.61	2010
Czech	<i>moc</i>	7.67	2149	<i>muset</i>	7.11	1229
Russian	<i>(s)moč'</i>	8.19	3594	<i>(dolžen)</i>	(6.99)	1091
Lithuanian	<i>(ne)galėti</i>	8.74	6234	<i>reikėti</i>	7.61	2015
Welsh	<i>gallu</i>	7.73	2267	<i>(rhaid)</i>	NOUN	
Estonian	<i>võima</i>	8.33	4156	<i>pidama</i>	(8.34)	(4183)
Finnish	<i>voida</i>	8.41	4479	<i>täytyä</i>	?	
Hungarian	<i>(tud 'know,can')</i>	7.90	2688	<i>kell</i>	8.06	3156

Table 8. The major modal verbs expressing Possibility and Necessity.⁴

⁴ Russian *smoč'* pf. 'become possible' 353/5.87 is included; *moč'* + *smoč'* 3,594/8.19 *dolžen* 'must' is an adjective; Welsh *rhaid* is a noun meaning 'necessity' (King 2016: 264); Estonian *pidama* represents two homonymous verbs meaning 'must' and 'hold' (Raag 2014: 317), which both appear to be frequent; The Finnish cognate *pitää* 'hold' can also express obligation. Hungarian *tud* means both 'know' and 'can'.

The identification of meanings in Table 8 is primarily based on the glosses in the frequency dictionaries. The detailed descriptions in Hansen & de Haan (2009) have also been consulted, although it has not been possible to account for this in detail. A further resource that has been used is the Word Alignment Database (WAD) included as a tool in OPUS (Tiedeman 2012). The analysis is representative also for analyses carried out on the nuclear verbs.

Even though MUST has been used to symbolize the major modal verb that expresses Necessity, *must* is rivalled by alternative expressions in English. As can be observed in Table 8, *must* has a lower frequency than its closest correspondences in the other European languages. This is the result of an ongoing change. Aijmer (2019) shows that *must* is the main proponent for epistemic modality in present-day English but is losing ground to *have to* and other semi-modals in the area of deontic modality and this development is related to a cultural change towards the use of less authoritative and face-threatening language that appears to have parallels in some other European languages. Variable expression of MUST is characteristic of further languages. In Finnish, a frequent alternative to express MUST is *täytyy* (3SG.PRES of *täytyä*), which is used with a subject in the genitive: *Minun täytyy lähteä* ‘I (lit. ‘my’) must leave’. In addition to *täytyy*, *pitää* ‘hold’ is a frequent translation of *must* in Subtitles. In Europarl, however, *on* ‘be.PRS.3SG’ dominates completely in a special construction with a passive present participle as in (2).

(2) Finnish (Europarl)

<i>taistelu on</i>	<i>käy-tä-vä</i>	<i>pääosin</i>	<i>kansainvälise-llä</i>	<i>taso-lla</i>
battle is	employ-PASS-PRS.PTCP	mainly	international-AD	level-AD

‘This battle must, in the main, be fought at international level.’

To sum up, CAN is realized as one of the 3-6 most frequent verbs in the majority of European languages and its frequency expressed as LN typically reaches above 8.0, which is superseded only by BE and HAVE. MUST is also expressed as a frequent verb and its LN usually reaches above 7.0, which is the cut-off point for the 20 most frequent verbs, but the picture is more complex than for CAN. One reason for that is that constructions with HAVE (or in Finnish BE) are frequent alternatives to a modal verb in several languages especially in certain text types.

5. MAKE and DO

Both *make* and *do* belong to the most frequent verbs in English, but in most European languages this verb pair has one major correspondence, which is symbolized by MAKE in Table 9. Dutch has been included, since it is one of the few European standard languages that have a do-verb. For a wide range of functions of DO in various varieties of Germanic languages, see Tieken-Boon van Ostade et al. (1998). See also Juvonen (2016) on the use of MAKE/DO as light verbs in pidgin lexica.

Language	MAKE				DO			
	<i>form</i>	<i>Rank</i>	<i>LN</i>	<i>Pm</i>	<i>form</i>	<i>LN</i>	<i>Rank</i>	<i>pm</i>
English	<i>make</i>	9	7.68	2173	<i>do</i>	8.63	4	5596
German	<i>machen</i>	9	7.59	1971	<i>tun</i>	6.62	21	752
Swedish	<i>göra</i>	9	7.89	2669				
French	<i>faire</i>	4	8.37	4330				
Spanish	<i>hacer</i>	5	8.31	4081				
Greek	<i>káno</i>	7	7.65	2101				
Czech	<i>dělat/udělat</i>	14	6.90	997				
Russian	<i>delat'/sdelat'</i>	13	7.40	1640				
Lithuanian*	<i>(pa)daryti</i>	22	7.25	1409				
Welsh	<i>gwneud</i>	3	8.46	4738				
Estonian	<i>tegema</i>	6	8.04	3114				
Finnish	<i>tehdä</i>	6	8.26	3896				
Hungarian**	<i>tesz</i>	6	7.31	1490				

*Lithuanian *sudaryti* 'form; conclude': 1266 pm

**Hungarian *tesz* 'make, do; put'

Table 9. MAKE and DO in European languages.

Gilquin & Viberg (2009) and Viberg (2006b) distinguish 6 main functions of *do* and *make* displayed in Table 10, which will be taken as a point-of-departure for the following brief survey. The function Pro-Verb includes two main subtypes.

Pro-verb covers what Van Valin & LaPolla (1997) refer to as a *Generalized Action Verb*, which is used when the nature of the event is unknown as in *I wonder what he did*. This function can be regarded as basic for languages that have a specific do-verb. Pro-verb also covers the anaphoric use as in: *He left and she **did** too*. Dutch *doen* is used as a pro-verb in both these uses, but this does not apply to German *machen*, which is

used as a generalized action verb: *Was macht er?* ‘What is he doing?’ The number of do-languages is too small to form general rules. Another case that would be worth looking at is the alternation between *lave* and *gøre* in Danish. The use of English *do* in interrogative and negative sentences stands for a large proportion of the number of tokens. It is unique at least among European languages and will not be commented on further (on *do* as a placeholder, see Hammarberg & Viberg 1977: 150–151).

Main functions	Examples
Production	She <i>made</i> her own wedding dress.
Transformation	It was this movie which <i>made</i> him a star.
Causative-ADJ	What <i>makes</i> you so happy?
Causative-VP	I like him because he <i>makes</i> me laugh.
Support verb	I can’t <i>make</i> a decision just yet.
Pro-verb:	
Generalized Action Verb	What did they <i>do</i> ?
Anaphoric verb	He left and she <i>did</i> too.

Table 10. Main functions of DO/MAKE.

For *make*, Production (or Creation), which refers to causing something to come into existence, can be regarded as the basic function and all the verbs symbolized as MAKE in Table 9 can be used to express this meaning. MAKE can be regarded as the nuclear verb of the semantic field, which also contains more specialized verbs such as *sew*, *build*, *manufacture* and *produce*. This would be worth a more extensive discussion since some of the other more abstract and grammatical meanings are more frequent (see Viberg 2006b for Swedish). The decisive factor is that production is a meaning from which the other meanings can be derived by extension in the most natural way. Transformation, which follows Production in Table 10, can be used with a construction typical of a verb of Production: *The director managed to make a star* [Product] *out of this unexperienced teenager* [Material] and the causative meanings Causative-VP and Causative-ADJ are more abstract extensions with a modified argument structure. In examples such as *make a decision* and *do harm*, the action is specified in the noun and *do* and *make* serve as support verbs. Both DO and MAKE have this function also in other languages, but this function will not be discussed further since the usage patterns vary a lot across languages with respect to which

abstract nouns are used. Table 11 shows the distribution of possible functions. (Data on English, Dutch, Swedish and French are based on Gilquin & Viberg 2009).

	ENG		DUT		SWE	FRE	GRE	RUS	FIN
Production	<i>make</i>		<i>maken</i>		<i>göra</i>	<i>faire</i>	<i>káno</i>	<i>delat'</i>	<i>tehdä</i>
Transformation									
Causative-ADJ						<i>rendre</i>			
Causative-VP				<i>doen</i>	<i>få</i>	<i>faire</i>			<i>saada</i>
Support verb		<i>do</i>	<i>maken</i>		<i>göra</i>				<i>tehdä</i>
Pro-verb									

Table 11 Possible functions of DO/MAKE in selected European languages.

Schultze-Berndt (2008) in a typological study of do-verbs⁵ based on a sample of nine very diverse non-European languages identifies further functions, some of which are characteristic also of European languages. The defining feature for a do-verb is that it can function as a Generalized Action Verb in the sense of Van Valin & LaPolla (1997). Like the European languages with a combined meaning DO/MAKE, languages like Kalam (kala1397; Nuclear Trans New Guinea, Madang) and Samoan (samo1305; Austronesian, Malayo-Polynesian) have a do-verb that also functions as a verb of Creation (Production). In some languages, do-verbs also function as productive causative markers, for example Ewe (ewee1241; Atlantic-Congo, Volta-Congo), but the presence of the creation sense seems to be a precondition for that use. (Dutch *doen* is an exception.) Crosslinguistically, a common use of do-verbs is to introduce quotations (Schultze-Berndt 2008: 191–192). European languages primarily rely on a specific verb SAY (see Section 8), but *faire* can be used to introduce direct reports in French, even though this use is not very frequent (Gilquin & Viberg 2009: 76). It is, however, very frequent in some regional varieties of spoken Italian (Francesca di Garbo, personal communication). A related function, which is widely attested crosslinguistically, is the use of a do-verb as a verbalizer with sound-symbolic elements. In several languages, the do-verb functions as an equivalent of ‘feel’. In

⁵ Schultze-Berndt’s term do-verb covers also the verbs referred to as MAKE in this paper. Her choice of term is obviously motivated by the focus on the use as generalized action verb and appears to be compatible with the analysis in the present paper with respect to the phenomena that are discussed in both papers.

Hausa, for example, *yi* ‘do’ is used to express emotions such as ‘be angry’, ‘be happy’ and ‘feel regret’.

6. Motion: GO, COME and the variable PUT

There are two basic types of motion verbs. *Come* and *go* refer to subject motion, where the moving entity functions as a subject, whereas verbs such as *put*, *pull* and *carry* refer to object motion (also referred to as caused motion). As shown in Table 12, verbs meaning GO and COME appear among the 20 most frequent verbs in all the studied languages. Among the object-motion verbs, the verbs of putting tend to be most frequent and come close to being nuclear verbs. For that reason, they will be discussed at some length in Section 6.3.

Lang.	GO			COME			PUT		
	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>
ENG	<i>go</i>	7.82	2495	<i>come</i>	7.33	1519	<i>put</i>	6.55	700
GER	<i>gehen</i>	7.41	1648	<i>kommen</i>	7.52	1837	POSTURAL		
SWE	<i>gå</i>	7.81	2476	<i>komma</i>	8.12	3348	POSTURAL		
FRE	<i>aller</i>	7.69	2194	<i>venir</i>	6.76	860	<i>mettre</i>	6.75	855
SPA	<i>ir</i>	8.17	3518	<i>venir</i>	6.68	799	<i>poner</i>	6.92	1017
GRE	<i>páo</i>	6.61	741	<i>érxume</i>	6.43	622	<i>θέτο</i>	6.47	647
							<i>vázo</i>	5.13	169
CZE	<i>jít</i>	7.07	1175	<i>přijít</i>	6.26	522	GIVE		
RUS	<i>idti/pojti</i>	7.19	1323	<i>prixodit'</i>	6.53	686	POSTURAL		
LITH	<i>eiti</i>	6.92	1011	<i>ateiti</i>	5.95	385	<i>dėti</i>		
WEL	<i>mynd</i>	8.22	3726	<i>dod</i>	8.03	3072	GIVE		
EST	<i>minema</i>	7.75	2317	<i>tulema</i>	8.09	3264	<i>panema</i>	6.98	1073
FIN	<i>mennä</i>	7.02	1118	<i>tulla</i>	8.38	4371	<i>panna</i>		
HUN	<i>megy</i>	6.56	704	<i>jön</i>	6.48	653	<i>(tesz)*</i>	7.31	1490

*Hungarian *tesz* is glossed ‘do, make; put’ and represented in Table 9.

Table 12. GO, COME and PUT in European languages.

6.1. Come and go

COME relates the motion to a place that can be implicit: the Deictic Center (DC). In the prototypical case, DC is the place of the speaker at the time of utterance: *Come (here)!* This serves as a defining feature. DC can be transferred step by step, but here

languages differ. DC can be transferred to the place of the speaker at reference time, to the place of the hearer, or even to the central character in a third person narrative. GO is in principle negatively defined and covers the cases not covered by COME in individual languages. Fillmore (1997) is the classical discussion of *come* and *go* in English. Ricca (1993) studied the contrast between COME and GO in 20 European languages. Lewandowski (2014: 46) proposed a universal hierarchy of extensions from the prototype with four steps, which is a good starting point for the discussion, since it emphasizes the great variation with respect to the identification of DC across languages. In Portuguese (step 1), COME is restricted to the speaker's location at speaking time but cannot be extended to speaker's location at reference time, whereas the latter is possible in Spanish (step 2). English (step 3) in addition allows the extension to the addressee's location, whereas in East and West Slavic languages (step 4), COME can refer to any Goal. The situation in languages representing the fourth step is studied more closely in Martinková & Janebová (2024), who look at the use of venitive verbs in Czech, where the closest correspondences of *come* as in other East and West Slavic languages are verbs with the perfectivizing prefix *při-*. This prefix productively combines with basically imperfective manner-of-motion verbs such as Czech *přijít* [*při-walk*], *přiběhnout* [*při-run*], *přijet* [*při-drive*] and *přilétet* [*při-fly*]. These verbs are not manner neutral like English *come*. Lithuanian is another language, where COME is not a simple verb but a morphologically marked version of GO: *eiti* 'go' – *ateiti* 'come'. Interestingly, German has a manner neutral verb *kommen* but also deictic particles *her-* (roughly 'to DC') and *hin-* ('from DC'), which can be combined with manner-of-motion verbs (see Meex 2020, for a more exact description). There are also great differences even between closely related languages such as English and Swedish with respect to the range of extended meanings. With respect to grammatical meanings, Swedish uses *komma* as a future marker, whereas English uses *go* (*be going to*). In addition, Swedish *gå* 'go' is used to express a specific modal meaning. There are also many abstract lexical meanings that contrast across the two languages (Viberg 1999a, 2003).

Even if the universality of the distinction between COME and GO was questioned already in a thought-provoking article by Wilkins & Hill (1995) and is clearly not an absolute universal, it represents a wide-spread lexical contrast. Based on a study of the Gospel According to Mark in 100 languages, Wälchli & Cysouw (2012) found that a distinction between COME and GO was made in most languages. This system is not always stable. DeLancey (1985) has shown how Tibeto-Burman languages change in a cyclical fashion. In essence, at a certain stage 'come' develops into a bound affix

which can be combined with other motion verbs, among them ‘go’. At a later stage, ‘come’ can arise anew through the merger of this affix with ‘go’. There are also languages that make further distinctions and use special verbs to indicate whether DC serves as the home base or not. In Diuxi Mixtec spoken in Oaxaca, Mexico (Kuiper & Merrifield 1975), there is a set of motion verbs which contrast with respect to both deictic distinctions and whether a location is the homebase (Base) of the subject of the verb, for example *ndisi* ‘move to Base toward DC’ and *vasi* ‘move to non-Base toward DC’. Deixis in motion verbs is dealt with systematically and in an innovative way in Matsumoto (2025).

6.2. Basic contrasts between European languages: Assisted and Indeterminate motion

There is a geographical belt in Europe, where motion in a vehicle is obligatorily marked in the verb (Viberg 2013b,c). Continental Scandinavian languages together with German and the Baltic and the East and West Slavic languages form a block and make a lexical distinction between moving on foot and travelling in a vehicle and do not have a general motion verb corresponding to English *go* (general ‘go’) but rather a verb which can be symbolized as ‘go = walk’. In Swedish, for example, the verb *åka* must be used in a sentence such as *Barnen åker buss till skolan* ‘The children go by bus to school’. *Barnen går till skolan* can only mean that they go by foot.

In East and West Slavic languages, there is also a lexical contrast between determinate and indeterminate motion verbs (Jakobson 1971, Sussex & Cubberley 2006: 482–483). Determinate verbs describe motion towards a Goal or Endpoint, whereas indeterminate motion verbs refer to aimless motion without a specific goal (‘go around’), habitual repetition and ability to walk, run etc. Not all motion verbs contrast in this way and the number of pairs vary in Slavic languages between 9 and 20.

6.3. Other basic motion verbs

PUT appears as one of the 20 most frequent verbs in several European languages, but Germanic and Slavic languages tend to make an obligatory differentiation between two or three verbs of putting based on postural verbs (SIT/STAND/LIE). As discussed in Viberg (2015a), the number of contrasts between the postural verbs have been reduced and in some languages one such verb has developed towards a general put verb. In addition, there is a general trend towards the development of a general verb

of putting with another basic meaning. In some languages, DO, LET or GIVE are used to fulfill that function. Lenz et al. (2020) describe the use of German *tun* ‘do’ as a general put verb in non-standard varieties across the whole German speaking area. This study also shows how *dát* ‘give’ in Czech is the most frequent put verb in the spoken language and the second most frequent (after *položít* ‘lay’) in the literary language. From a typological perspective, verbs of putting have been studied in Kopecka & Narashimhan (2012).

There are no basic motion verbs other than COME and GO that appear among the 100 most frequent verbs in all 13 languages. There are, however, some that appear in most of them. A verb glossed LEAVE appears in 11 languages and 10 of the 13 languages have a verb glossed FOLLOW, which refers to co-motion. The basic meaning of FOLLOW is ‘move behind and in the same direction as some other mover’ even though the meaning may be wider in various ways. In Swedish, *följa* corresponds both to FOLLOW and to ACCOMPANY ‘move together with’. The two meanings can be kept apart by adding a particle: *följa efter* (‘after’) ‘follow’, *följa med* (‘with’) ‘accompany’.

With respect to possible language-specific features, it is interesting to look for possible reflections of Talmy’s (1985) typology of motion events in the lists of the 100 most frequent verbs. Verbs referring to motion into and out of a container (ENTER/EXIT) are found in this frequency range in the Verb-framed languages that typically encode Path in the verb (rank within parentheses): French *entrer* (88), Spanish *entrar* (41) and Greek *béno* (72), whereas French *sortir* (77), Spanish *salir* (28) and Greek *vjéno* (100) express EXIT. Among European languages, the use of frequent verbs meaning ENTER and EXIT is a characteristic feature of Verb-framed languages. The situation is less clear with respect to verbs referring to motion UP/DOWN. Besides verbs glossed ASCEND and DESCEND, which are characteristic of Verb-framed languages, certain such verbs such as FALL appear among the 100 most frequent verbs in 4 languages (ENG, GER, HUN, SPA) and in Swedish *falla* has rank 103. Verbs covering the concepts of RISE, RAISE and/or LIFT appear among the 100 most frequent verbs in 6 languages (FRE, SPA, EST, FIN, LIT, WEL).

7. Possession: GIVE and TAKE versus TAKE and BRING

Traditionally, TAKE is regarded as the opposite of GIVE, but in studies of English, *take* and *bring* have also been regarded as transitive or causative counterparts of *go* and

come, since deixis is also involved in the contrast (Fillmore 1997, Levin 1993). Both perspectives are valid.

7.1. GIVE and TAKE

The data for the nuclear verbs GIVE and TAKE are shown in Table 13.

Language	GIVE			TAKE		
	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>
English	<i>give</i>	7.18	1314	<i>take</i>	7.49	1792
German	<i>geben</i>	7.71	2240	<i>nehmen</i>	6.43	618
Swedish	<i>ge</i>	7.24	1399	<i>ta</i>	7.69	2189
French	<i>donner</i>	6.75	851	<i>prendre</i>	6.96	1058
Spanish	<i>dar</i>	7.61	2012	<i>tomar</i>	6.45	632
Greek	<i>δίνο</i>	7.43	1689	<i>πέρο</i>	6.86	950
Czech	<i>dát/dávat</i>	7.01	1111	<i>vzít/brát</i>	6.18	483
Russian	<i>davat’/ dat’</i>	6.86	956	<i>vzjat’/brat’</i>	6.54	694
Lithuanian	<i>duoti</i>	5.88	357	<i>(pa)imti</i>	6.90	991
Welsh	<i>rhoi</i>	7.57	1942	<i>cymryd</i>	6.46	638
Estonian	<i>andma</i>	7.52	1851	<i>võtma</i>	7.54	1884
Finnish	<i>antaa</i>	7.35	1550	<i>ottaa</i>	7.37	1581
Hungarian	<i>ad</i>	6.94	1033	<i>vesz</i>	7.05	1156

Table 13. The frequencies of GIVE and TAKE.

There is an extensive literature on GIVE from a crosslinguistic perspective (Newman 1998, Nolan et al. 2015, Bouveret 2021). The seminal work is Newman (1996), who presents an extensive analysis of the basic meaning of English *give* and its rich semantic network from a cognitive linguistic perspective in combination with a broad typological survey. Across languages, literal GIVE tends to extend to more abstract fields, in particular Interpersonal communication (*give advice to someone*), and to Emergence (of Sound, for example: *give a yell*). Newman also discusses the development of grammatical uses such as Enablement and Causation as well as Recipient and Benefactive markers.

The status of GIVE as a nuclear verb appears to represent a strong universal tendency, but it is clearly not an absolute universal. In some languages, such as Sango (sang1328; Atlantic-Congo, Volta-Congo), GIVE and TAKE are the same verb and the semantic distinction between them is signaled by the construction (Thornell 2002).

Neither is GIVE always a simple verb. Reesink (2013) describes Papuan languages where GIVE is expressed with a combination of serial verbs and Rice (1998) shows how Chipewyan (chip1261; Athabaskan-Eyak-Tlingit, Athabaskan) completely lacks specialized verbs of possession and uses physical action verbs describing handling and manipulation as the closest equivalents of the verbs of possession. In Japanese, GIVE is expressed with a set of verbs that are differentiated with respect to the social status and interpersonal relations between speaker and hearer and the giver and the receiver (Loveday 1986: Chapter 3).

Viberg (2010, 2025) shows how TAKE as a verb of Possession can be analyzed as a goal-directed physical action verb and how an important concrete use of TAKE can be described as extending the arm and grasping an object as the first part of a chain of actions in order to gain control i.e. possession of the object, see (3). Ownership is often not an issue in this use, only physical possession.

		VERB
(3)	And he took the coarse chunks of bread, one in each hand, and ate.	<i>take</i>
FRE	et prit les morceaux de pain noir,	(source: MPC) <i>prendre</i>

The verbs used in this example are the ones identified as TAKE in Table 13. In addition to this concrete use, TAKE has many abstract extended uses.

7.2. TAKE and BRING

In many languages, another verb than TAKE is preferred (or required) when the verb refers to caused motion of the grammatical object combined with the co-motion of the Agent as in (4). The meaning of these verbs will be referred to as TAKE^{CM} (CM: Co-Motion).

		VERB
(4)	She took me to see this movie.	<i>take</i>
FRE	Elle m' a emmené voir le film.	<i>emmener</i>
	she me has taken ^{CM} see the film	(source: OPUS Subtitles)

In English, *take* and *bring* are regarded as opposites, but in languages that use separate verbs to express TAKE and TAKE^{CM}, it is the verb with the latter meaning that is the opposite of BRING. TAKE^{CM} refers to motion from DC and contrasts with BRING, which refers to motion to DC. As is the case with COME, the rules for assigning DC varies across languages. BRING is represented by a frequent special verb in all the studied

languages except Welsh and, in actual practice also Swedish, in which *bringa* has rank 933 and a frequency of only 14 pm. BRING is characteristic of European languages, even if it is not as frequent as any of the nuclear verbs.

In French, there are two pairs of verbs that correspond to TAKE^{CM} and the same applies to BRING. The pair *emporter/emmener* refer to motion from a deictic center (DC) and correspond to TAKE^{CM}, whereas *apporter/amener* refer to motion to DC and correspond to BRING. In principle, the first verb in each pair is used with inanimate objects and the second with human objects. The verbs are derived versions of *porter* ‘carry’ and *mener* ‘lead’. Verbs meaning CARRY and LEAD also describe co-motion but do not refer to DC. The contrasts between TAKE, TAKE^{CM} and BRING are summed up in a selection of representative European languages in Table 14.

Language	TAKE ^{CM}			BRING		
	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>
English	<i>take</i>	See Table 13		<i>bring</i>	6.08	439
Swedish	<i>ta</i>			<i>(bringa)</i>	2.64	14
French	<i>emporter</i>	4.01	55	<i>apporter</i>	5.06	158
	<i>emmener</i>	3.76	43	<i>amener</i>	4.44	85
Spanish	<i>llevar</i>	6.75	853	<i>traer</i>	5.52	250
Greek	<i>páo</i> ‘go; cause to go’	See GO		<i>féro</i>	6.01	408
Welsh	<i>mynd âh</i> ‘go with’			<i>dod âh</i> ‘come with’		
Finnish	<i>viedä</i>	6.53	687	<i>tuoda</i>	6.73	841

Table 14. The frequency of TAKE^{CM} and BRING in some European languages.

In Greek, *páo* ‘go’ can be used as a co-motion verb and is a frequent translation of *take*, when it refers to co-motion as in (5).

(5) Greek (OPUS Subtitles)

θα σε πάο σπίτι με το aftokínitó μου
 FUT you.ACC go.1SG house with DEF car my
 ‘I’ll **take** you home in my car.’

In Swedish, *bringa* has a very low frequency and is used primarily with abstract objects: *bringa lycka* ‘bring luck’, *bringa klarhet* ‘throw light on’, *bringa ordning*, ‘bring order’. When *bring* has a concrete meaning, a frequent translation is *komma med* ‘come with’ but also *ta med* ‘take with’ and *ha med* ‘have with’ are frequent. (The situation is similar in Norwegian, see Ebeling 2017). In Welsh, both co-motion verbs are rendered

with more analytical expressions: TAKE^{CM} is expressed with *mynd âh* ‘go with’ and BRING with *dod âh* ‘come with’ King (2016: 469, 471). In Czech, Russian and Lithuanian, TAKE^{CM} has three major correspondences (see, for example, Naughton & von Kunes 2021: 214 for Czech.). BRING is expressed by adding a prefix to these verbs. TAKE and BRING in European languages are treated more in detail in Viberg 2025 ms.) There is an interesting typological study by Margetts et al. (2022) of caused accompanied motion (my TAKE^{CM} and BRING) based on data from eight languages from the Americas and from the Pacific area. The study shows how caused accompanied motion is described in a more analytical way, for example by transitivizing COME and GO by adding causative or applicative markers (see also Margetts et al. 2022).

7.3. *Summing up Verbs of Possession*

GIVE serves as a nuclear verb in European languages and tends to be basic even in languages outside Europe, but this is only a tendential universal. As a verb of possession and as an antonym of GIVE, TAKE has a similar status although its use in non-European languages is not as well documented as that of GIVE. The use of TAKE as a co-motion verb (TAKE^{CM}) and as an antonym of BRING should be regarded as a language-specific characteristic of languages such as English and Swedish. In these languages TAKE can be extended to cover TAKE^{CM}, whereas this meaning is expressed with other verbs than TAKE in most European languages. BRING is characteristic of European languages except some like Swedish, Norwegian and Welsh, where this meaning is expressed analytically by COME + WITH. Verbs glossed RECEIVE and GET, which both appear among the 100 most frequent verbs, will be commented on in Section 11.3. A couple of other possession verbs tend to be frequent but do not appear among the 100 most frequent verbs in all 13 languages. PAY appears within this range in 9 languages and BUY in 7.

8. Verbal Communication: SAY and TELL

Table 15 shows the frequency of SAY, which is the most frequent verbal communication verb in European languages.

A very large inventory of Verbal Communication Verbs (VCVs) has been described for several European languages. Ballmer & Brennenstuhl (1981) account for 4,800 English speech act verbs (including multiword expressions) and Harras et al. (2004,

2007) devote two volumes to German VCVs. Wierzbicka's (1987) analysis of English VCVs is particularly relevant for the present study, since the primitive SAY appears in the definitions of all the other VCVs, which can be interpreted as support for the idea that SAY is a nuclear verb.

Language	SAY	LN	pm
	<i>form</i>		
English	<i>say</i>	8.11	3335
German	<i>sagen</i>	7.99	2954
Swedish	<i>säga</i>	7.96	2868
French	<i>dire</i>	7.97	2898
Spanish	<i>decir</i>	8.29	3967
Greek	<i>léyo</i>	8.07	3183
Czech	<i>řict/řikat</i>	7.72	2249
Russian	<i>skazat'/govorit'</i>	8.16	3503
Lithuanian	<i>(pa)sakyti</i>	7.72	2260
Welsh	<i>dweud</i>	7.74	2308
Estonian	<i>üttelema</i>	7.83	2523
Finnish	<i>sanoa</i>	7.98	2913
Hungarian	<i>mond</i>	7.72	2262

Table 15. The nuclear Verbal communication verb SAY.

8.1. The nuclear verb SAY

SAY appears to have a correspondence in most (possibly all) languages outside Europe, although SAY also has a tendency to acquire a wide range of more general meanings. Goddard (2001: 31–32) in a survey of lexico-semantic universals states: “In many languages there is a unique lexical form meaning ‘say’, as in English. But there are languages in which the word for ‘say’ can also express the meanings ‘do’ (as in Samoan) or ‘want’ (as in Kalam and Mangaaba-Mbula). Another common polysemy in which ‘say’ is involved concerns ‘making sounds’ of various kinds.” In Papuan languages, SAY according to Foley (1986) is the verb most commonly used in periphrastic causative constructions and it also tends to be used in desiderative constructions. Another geographical area where SAY has developed many secondary meanings is Ethiopia, where one of the most characteristic secondary meanings is

intention (Ferguson 1976: 71). In addition, SAY has a strong tendency to grammaticalize into a quotative or a more general complementizer (Lord 1993: Chapter 7, Heine & Kuteva 2002).

8.2. Other basic verbs of verbal communication

SAY and TELL. English *tell* has rank 22 and the frequency 772 pm (LN 6.65). Its frequency is only slightly lower than what is characteristic of a nuclear verb. *Tell* must, however, be regarded as a language-specific basic verb, since the closest correspondences in the other European languages tend to have a much lower frequency and the most frequent VCVs, which are identified as SAY in this study, are glossed ‘say, tell’ in some of the frequency lists. As discussed in a corpus-based contrastive study of *tell* and its Swedish correspondences (Viberg 2017b), *tell* can be used as an information verb and report a statement. In that function, *tell* is often translated into Swedish by *säga* ‘say’. The verbs *si* ‘say’ in Norwegian and *dire* ‘say’ in French are also frequent correspondences of *tell* (Egan 2021). A distinctive characteristic of *tell* as an information verb is that the Addressee is explicitly mentioned in most cases (as much as 90% in Viberg 2017b: 64). The semantically closest equivalent to *tell* can usually serve as a verb that reports a narrative, which typically refers to a complex set of events that may be fictional. These correspondences have different ranges of uses in addition to the prototypical ‘tell a story’ (see Viberg 2017c).

SAY and SPEAK. In English, the verb *say* typically requires explicit reference to the Message (*He said that he would leave / He said so*) but can optionally refer also to the Addressee and the Topic: *He [Speaker] said something [Message] about it [Topic] to Alice [Addressee]*. Speech activity verbs refer to the act of communicating verbally without referring to the message. They can refer simply to the activity as in *He was talking* but can optionally include reference to the Addressee and the Topic: *He [Speaker] talked to Alice [Addressee] about the wedding [Topic]*. European languages usually have one or sometimes like English (*speak, talk*) two speech activity verbs that belong to the 100 most frequent verbs. At least one speech activity verb appears among the 100 most frequent verbs in 12 of the 13 languages in the sample. The exception is Russian in which the situation is more complicated. SAY is expressed by two suppletive aspect forms *skazat’* pf and *govorit’* impf., but in particular the imperfective form is used also as a speech activity verb.

WRITE. A verb meaning ‘write’ appears among the 100 most frequent verbs in all 13 languages. In English, this verb can be used both as a reportative verb *He wrote that he would come* and as an activity verb: *He has been writing all day. He wrote a few words [Message] about it [Topic] to me [Addressee] in a letter [Medium]*.

ASK and ANSWER are also among the most frequent VCVs. In English, *ask* can be used both in the sense ‘ask-inquire’ (*He asked what I was doing*) and in the sense ‘ask-request’ (*He asked me to open the window*). All 13 languages have a verb that can mean ask-inquire among the 100 most frequent verbs (including some cases in which this verb can also mean ‘ask-request’ such as English *ask* and French *demander*). ANSWER appears among the 100 most frequent verbs in 7 languages. The polysemy of *ask* and *answer* is described in Rudzka-Ostyn (1989, 1995) and contrastively compared with the Swedish correspondences in Viberg (2016b).

9. Mental: SEE, KNOW and WANT

Mental is a macro-field that comprises several general semantic fields: Perception, Cognition, Emotion and Desire.

9.1. Perception

Verbs of perception can refer to intentional Activities such as *look* and *listen* or to non-agentive Experiences such as *see* and *hear*. Only Experiences will be discussed here. (See Viberg 2019 for the general structure of the semantic field of perception verbs.) Table 16 shows the frequencies of the verbs meaning SEE.

SEE is the most frequent perception verb in all European languages and with a few exceptions has an LN above 7, which is characteristic of the nuclear verbs, whereas HEAR and FEEL have a lower frequency. LN is usually above 6 for HEAR but not above 7. FEEL is more variable and ranges between 5.4 – 7.0 (214 – 1067 pm). The rank order between the latter two verbs varies across European languages. The major reason for this appears to be that the verbs identified as FEEL in some languages have very frequent cognitive meanings (see Section 9.2.1). Verbs referring to Experiences of TASTE and SMELL have a much lower frequency.

SEE appears to be the most basic perception verb also in non-European languages (Viberg 1983a). See also Evans & Wilkins (2000), San Roque et al. (2015) and Norcliffe & Majid (2024a,b). The status of SEE as a nuclear verb thus represents a very strong tendency across the world’s languages but is not an absolute universal. As discussed in

Viberg (2001), there are languages where SEE is not a simple verb as in Ngarinjin (non-Pama-Nyungan, Northern Kimberley, Australia).

Language	SEE	LN	Pm
	<i>form</i>		
English	<i>see</i>	7.6	1917
German	<i>sehen</i>	7.2	1357
Swedish	<i>se</i>	7.9	2592
French	<i>voir</i>	7.2	1344
Spanish	<i>ver</i>	7.3	2293
Greek	<i>vlépo</i>	6.9	1021
Czech	<i>(u)vidět</i>	6.6	712
Russian	<i>(u)videt'</i>	7.1	1227
Lithuanian	<i>(pa)matyti</i>	7.0	1039
Welsh	<i>gweld</i>	7.8	2460
Estonian	<i>nägema</i>	7.3	1542
Finnish	<i>nähdä</i>	6.9	1014
Hungarian	<i>lát</i>	7.0	1052

Table 16. The frequencies of SEE across European languages.

There are also languages that do not have any verb clearly tied to a specific sense modality. A well-documented case of this type is the Papuan language Kalam (Pawley 2020) from which example (6) is taken.

(6) Kalam (Pawley 2020: 290)

mñi nŋ-sp-an?

now perceive-PROG-2SG

‘Now do you notice/see/hear/smell/feel/understand (it)?’

9.2. Cognition

Cognition covers Knowledge, Belief, Understanding, Memory, Judgment, Desire, Intention, Problem solving, Planning and Resolution. (The concept Meaning is regarded as part of Verbal communication.) Verbs of thinking and other cognitive verbs have been discussed from a typological perspective in Fortescue (2001) and in Palmer et al. (2003), see in particular Goddard (2003) in that thematic issue.

9.2.1. KNOW

Etymology is particularly informative with respect to the verbs of knowing in European languages. According to Buck (1949: KNOW), the English verb *know* covers two meanings that were originally expressed by two different verbs and still is in many Indo-European languages. The first meaning is ‘know as a fact’ and verbs with that meaning are shown under KNOW in Table 17. The other meaning according to Buck is ‘be acquainted with (a person or thing)’. The meaning of these verbs, which are represented as KNOW2 in Table 17, is variable and it is difficult to pin down one basic meaning shared across all languages by these verbs.

Language	KNOW			KNOW2		
	<i>form</i>	<i>LN</i>	<i>pm</i>	<i>form</i>	<i>LN</i>	<i>pm</i>
English	<i>know</i>	7.5	1855	---		
German	<i>wissen</i>	7.2	1358	<i>kennen</i>	6.0	400
Swedish	<i>veta</i>	6.9	1032	<i>känna</i> ‘feel’	7.0	1067
French	<i>savoir</i>	7.3	1423	<i>connaître</i>	6.1	461
Spanish	<i>saber</i>	7.5	1855	<i>conocer</i>	6.4	613
Greek	<i>kséro</i>	6.3	547	<i>γνóρίζο</i>	6.1	448
Czech	<i>vědět</i>	7.4	1691	<i>znát (se)</i>	5.4	220
Russian	<i>znat’</i>	7.4	1579	<i>znat’</i>	= KNOW	
Lithuanian	<i>(ne)žinoti</i>	7.37	1589	<i>pažinti</i>		
Welsh	<i>gwybod</i>	6.9	961	<i>adnabod</i>		
Estonian	<i>teadma</i>	7.6	2058	<i>tundma</i> ‘feel’	6.5	997
Finnish	<i>tietää</i>	6.8	873	<i>tuntea</i> ‘feel’	6.5	639
Hungarian	<i>tud</i> ‘know, can’	7.9	2688	<i>ismer</i>	5.8	332

Table 17. Verbs of knowledge in European languages.

Latin *sapere* ‘taste; have sense, be wise’ is the source of KNOW in the Romance languages except Romanian: French *savoir*, Spanish *saber* and Italian *sapere*; Romanian *ști* (from Latin *scire*). In most Germanic and Slavic languages, KNOW can be traced back to the Indo-European root **weid-*, perfective **woida* (have seen > know) and this represents the most well-known example of the extension from SEE to KNOW (see Sweetser 1990). Examples are Swedish *veta*, German *wissen*, Czech *vědět* and Polish *wiedziéć*. Another source is IE **gēnə-/gēnō-*. This is the source of English *know* and Russian *znat’* but also of a second verb (KNOW2) in several languages such as French *connaître* and Spanish *conocer*.

and Greek *gnorízo*. The same root has also developed into Gothic *kunnan*, which has developed further into a modal verb. The German verb *kennen* and Swedish *känna* are causative versions (Gothic *kan+jan*), that represent KNOW2. In the Fenno-Ugrian languages, Hungarian *tud* ‘know, can’ is historically related to Estonian *tundma* and Finnish *tuntea* (Collinder 1955: 63), which both cover ‘feel’ as well as KNOW2. The semantic distinction in Greek between *kséro* KNOW and *gnorízo* KNOW2 is unclear. Both verbs can refer to ‘know as a fact’. Variation across registers appears to be involved. *Kséro* is the dominant translation of *know* in Subtitles, whereas *gnorízo* dominates in Europarl based on WAD.⁶

Table 18 shows how some representative meanings of *know* are distributed in a selection of European languages (see also Viberg 2005, 2015b).

Meaning	Example	SWE	SPA	CZE	RUS	LIT
KNOW A FACT	Mary knows that Harry left	<i>veta</i>	<i>saber</i>	<i>vědět</i>	<i>znat’</i>	<i>(ne)žinoti</i>
KNOW ABOUT A FACT	Mary didn’t know of/about the decision	<i>känna</i> <i>till</i>	<i>conocer</i>	<i>znát</i>	<i>znat’</i>	?
KNOW A PERSON	Mary knows Harry	<i>känna</i>	<i>conocer</i>	<i>znát</i>	<i>znat’</i>	<i>pažinti</i>
RECOGNITION	Mary recognized Harry	<i>känna</i> <i>igen</i>	<i>reconocer</i>	<i>poznat</i>	<i>uznavat’</i>	<i>(ne)atpažinti</i>

Table 18. Distributions of meanings across KNOW and KNOW2.

Like English, Russian primarily uses one verb *znat’* to cover both KNOW and KNOW2, whereas the other languages use two verbs. A frequent meaning of Spanish *conocer*, which is not shown, is ‘meet’.

9.2.2. Verbs of thinking

The verb *think* equals the nuclear verbs in English in terms of its general meaning and its high frequency (R 15, 1539 pm, LN 7.3). It is, however, problematic to regard THINK as a nuclear verb characteristic of European languages in general because English *think*

⁶ Subtitles Δεν ξέρω πώς βρέθηκε στην εφημερίδα. <den *kséro* pós vréthike stin efemerída>
‘I don’t know how it got in the paper.’

EuroParl Όλοι γνωρίζουμε ότι η αιχμή της ανάπτυξης είναι μια διαδικασία γνώσης.
<óli *gnorízume* óti i aifóros anáptiksi íne mia diadikasía gnósis>
‘We all know that sustainable development is a learning process.’

appears to be much more general than the closest correspondences in many other European languages. A verb roughly glossed BELIEVE tends to come closer in frequency to THINK than in English, and in Swedish, *think* has three frequent translations (Viberg 2005: 142–143. See also the experimental study by Frommherz 2022).

The picture is complicated also by the fact that the verbs of thinking tend to develop special discourse and interactional functions, when they are used with first person present reference. Such uses can reach a very high frequency in some languages – especially in various types of spoken discourse. Aijmer (2015: 204) discusses the use of English *I think* as a discourse particle with very flexible meaning. Fetzer & Johansson (2012) compare the use with first person reference of the English verbs *think* and *believe* with French *penser* and *croire* in argumentative discourse. Parenthetical uses are most frequent with *I think* in English, whereas such uses are most frequent with *je crois* in French. Pragmatically based uses also are at play in Spanish and motivate why *creer* is more frequent than *pensar*, which is categorized as THINK. In Spanish according to Mulder (2017), *creo* ‘I believe’ stands out as marker of evidentiality and epistemic modality, whereas *pienso* ‘I think’ more often appears in its primary meaning and refers to the act of thinking or reasoning or is used to express intentions or plans.

9.3. *Desire*

Desire refers to verbs such as *want*, *desire*, *hope* and *will*. The nuclear verb in present-day English is *want* even if *will* can still have a similar meaning besides its use as future and modal auxiliary. The verb *want* can refer to a person who wants an object (*I want a beer*), someone who wants to do something (*I want to leave*) or someone who wants another person to do something (*I want you to leave*). *Want* can also refer to someone wanting other types of events to occur (*I want it to stop snowing*). European languages contrast with respect to which types of constructions are used to express such functions. The major correspondences of *want* are shown in Table 19.

In European languages, the situation where someone wants to do something is in general expressed with the construction NP₁ + VP^{Infinitive} (see (7)). In this construction, the subject is coreferential with the Wanter and left implicit in the complement. The only exception is Greek and some other Balkan languages, where the coreferential subject is marked by a person ending on the verb in the clausal complement (SBJV: subjunctive particle; *ði* [ðɛi] 3SG simple subjunctive, a perfective form).

(7) Greek (OPUS Subtitles)

kápios *théli* *na* *me* *ði*
 someone want.PRS.3SG SUBJUNCTIVE me see.PRF.3SG
 ‘Oh, a man just wants to see me.’

Haspelmath (2013) in a typological survey based on 283 languages shows that the complement subject is implicit in the majority of the languages (51%) and that an explicit subject is found in 25% of the languages. The rest signal the desiderative meaning with an affix or a particle. In some languages, a verb is used that also means ‘say’ and/or ‘think’. The typological study by Harkins (1995) based on 50 diverse languages gives several examples of that.

Language	WANT	LN	pm
	form		
English	<i>want</i>	6.85	943
German	<i>wollen</i>	7.61	2017
Swedish	<i>vilja</i>	7.34	1536
French	<i>vouloir</i>	7.37	1586
Spanish	<i>querer</i>	7.27	1435
Greek	<i>thélo</i>	7.12	1236
Czech	<i>chtít</i>	7.02	1120
Russian	<i>(za)xotet’</i>	7.07	1174
Lithuanian	<i>(ne)norėti</i>	7.40	1641
Welsh	<i>eisiau</i>	6.23	509
Estonian	<i>tahtma</i>	7.26	1427
Finnish	<i>halua</i>	7.32	1508
Hungarian	<i>akar</i>	6.77	871

Table 19. WANT in European languages.

9.4. Other basic verbs of cognition

Two more verbs in addition to the ones discussed above tend to appear among the 100 most frequent verbs. A verb meaning DECIDE can be found in 11 of the languages and a verb meaning UNDERSTAND in 10. Memory is an important part of cognition, but REMEMBER and FORGET appear among the 100 most frequent verbs only in a few of the languages.

10. How (un)usual are European languages? Universal and language-specific aspects of the inventories of the nuclear verbs in European languages

The characterization of basic verbs presented in this paper is intended to serve as a reference point for studies aiming to produce descriptions of basic verbs in the world’s languages. To what extent is it possible to find similarities and to what extent is it possible to identify languages that represent completely different types? More specifically, to what extent can equivalents to the nuclear verbs be found among the most frequent verbs in non-European languages and is it possible to identify other lexical verbs that hold a similar position?

The nuclear verbs do not represent any absolute universals in the sense that a direct equivalent can be found in all languages. In each of Sections 5-9, examples have been given of languages that lack a simple verb that expresses the corresponding meaning. It appears, however, that the nuclear verbs have frequent equivalents in many languages. Their status as universal tendencies appears to be an interesting working hypothesis that is worth testing. Viberg (2012) looked at data on the 20 most frequent verbs in some non-European languages with a special focus on nuclear verbs. As can be observed in Table 20, the nuclear verbs tend to be basic also in these languages.

Motion	Possession	Production	Verbal	Perception	Cognition	Desire
GO (9)	GIVE (6)	MAKE (9)	SAY (9)	SEE (9)	KNOW (8)	WANT (7)
COME (8)	TAKE (3)					
	GIVE/TAKE (1)					

Languages: Kurdish (Kurmanji), Turkish, Mongolian, Megrelian, Arabic, Chinese, Sango, Swahili, Lao.

Table 20. The nuclear verbs in 9 non-European languages.

The result supports the working hypothesis, but the test is based on a very restricted database and frequency data from a representative sample of the world’s languages will probably not be available in the recent future. A more promising approach would be to look at descriptions of languages with a restricted set of simple verbs. The most well-known are north Australian languages (more exactly languages in north-western and north-central Australia), see Dixon (1980: 280–281). An example that shows how such a language can look is presented in Table 21 based on the sketch of Nungali (nung1291; Mirndi, Yirram) in Hoddinot & Kofod (1976).

General Dynamic	Motion Subject	Motion Object	Possession	Physical Contact	Production
<i>-nand(u)-</i> 'be/stay'	<i>-Nga-</i> 'go'	<i>-ra-</i> 'put'	<i>-Nana-</i> 'give'	<i>-NaNga-</i> 'hit'	<i>-malin-ma-</i> 'make'
	<i>-ruma-</i> 'come'	<i>-djama-</i> 'bring'	<i>-ndja-</i> 'take'	<i>-maNu-</i> 'hit/kill'	
	<i>-da-</i> 'fall'	<i>-mili-</i> 'gather'	<i>-yal-gara-</i> 'send'		
	<i>-wada-gara-</i> 'follow'	<i>-Ngu-</i> 'throw'*			
Verbal Communication	Perception	Disconnection	Ingestion	Cooking	Other Fields
<i>-Ngu-</i> 'say (throw) '	<i>-Nawu-</i> 'see'	<i>-wiri-</i> 'bite'	<i>-ndji-</i> 'eat'	<i>-riga-</i> 'cook/burn'	<i>-yada-</i> 'spear'
	<i>-mala-Nawu-</i> 'hear'				<i>-nama-</i> 'dance'

*-Ngu- 'say/throw' is listed twice in accordance with its two meanings

Table 21. The 22 simple verbs in Nungali (Hoddinott & Kofod 1976).

Seven of the nuclear verbs have parallels according to the glosses: GO, COME, GIVE, TAKE, MAKE, SAY and SEE. HIT and EAT represent verbs that tend to have a high frequency in several non-European languages.

There is a systematic overview by McGregor (2002) of Australian languages with restricted sets of simple verbs according to which such languages have a very small inventory of inflecting verbs (from half a dozen up to around one hundred). These are typically combined with large sets of particles or other types of adjuncts to form compound verbs. Among the inflecting verbs, there are fourteen basic verbal meanings that tend to appear across languages with small sets of inflecting verbs. The lexical meanings of these verbs when they are used as simple verbs tend to belong to seven semantic domains characterized in the following way: "speech ('say, do'); stance ('sit, be', 'stand'); motion ('go', 'fall', and 'become'); induced motion ('carry', 'throw', 'put'); acquisition ('catch, get', 'give'); violence ('hit', 'poke'); and perception ('see')" (McGregor 2002: 153). This list includes approximate parallels to four nuclear verbs: SAY, GO, GIVE and SEE. HIT is a potential nuclear verb in the sense that it serves as the core within the field of physical contact verbs and has rich pattern of polysemy in many languages including grammatical uses (Viberg 2016c), but it does not appear among the 20 most frequent verbs in any European language. EAT has a wide range

of meanings in many languages including grammatical uses in some languages (Newman 2009).

A small, closed set of simple verbs is also characteristic of several Papuan languages. Foley (1986: 127) refers to “the comparative dearth of verb-stems” in many Papuan languages. This is in general compensated for in one of two ways, either by verb serialization, which is prominent among Sepik languages (e.g., Yimas, Alamlak, Barai), or by adjunct constructions, i.e. compound verbs consisting of nominal + verb-stem, which are prominent among highlands languages (e.g., Hua, Enga and Kewa). The most reduced number of verbs is found in the two closely related Kobon (kobo1249; Nuclear Trans New Guinea, Madang) (Davies 1981) and Kalam spoken in the Kaironk valley. Kalam has been described in quite some detail by Pawley (e.g., 1993), who has also written a survey article comparing Australian and Papuan languages with closed verb classes (Pawley 2006). There are only around 130 verb roots in Kalam, and among these the ten most frequent ones make up almost 80% of the verb roots in running text. Kalam can form complex verbs by adding adjuncts to the simple verbs as in north Australian languages but in addition uses verb serialization to a certain extent. The verbs that are most productively used to form complex verbs are glossed SAY, HAPPEN/MAKE, GO, COME and STABILIZE (PUT/FORM) (Pawley 2006: 1–6). Pawley also compares the meaning of the most frequent verbs in English with the most frequent verbs in Kalam and in Jaminjung, a north Australian language, whose system of simple and complex verbs has been thoroughly described in Schultze-Berndt (2000). Pawley’s basic conclusion is that the three languages are “not so different after all”, which lends some support to the claim that the meaning of the European basic verbs reflect universal tendencies. One general characteristic is that the languages with closed verb classes can have certain verb roots that cover the meaning of several of the European nuclear verbs. To what extent these “generalized nuclear verbs”, as they may be called, are polysemous and have separate senses corresponding to the various nuclear verbs or whether they have a more general meaning appears to be difficult to determine. An example of this is Kalam *ɲɲ*-‘perceive’, which (often as part of a complex verb) covers SEE in addition to other verbs of perception and also cognitive verbs such as KNOW and THINK. These meanings are closely related also in languages with open verb classes.

Many Indo-Iranian languages also have relatively small classes of simple verbs. Liljegren (2010: 55) shows how Palula (phal1254; Indo-European, Indo-Aryan) spoken in northern Pakistan places itself between the European languages and Kalam with respect to its reliance on its most frequent verbs. In Palula, the five most frequent

verbs account for 51.5% of the total textual occurrences of verbs and the 20 most frequent verbs account for close to 80%. Five of these correspond to nuclear verbs: MAKE, GO, COME, GIVE, SAY; In addition, HIT and EAT appear among the 20 most frequent verbs. There is a relationship between the size of the set of simple verbs and the morpho-syntactic structure of the language. Palula uses compound verbs consisting of V preceded by Noun or another Non-Verb Component: NVC + V to a greater extent than European languages, whereas Kalam relies extensively on both serial and compound verbs.

Persian according to Miller & Aghajanian-Stewart (2017) has around 150 simple verbs with any high frequency and according to Windfuhr & Perry (2014), around 18 of these are used as light verbs to form compound verbs together with a noun or other NVC, for example *goš dadæn* ‘listen’ (lit. ‘ear give’) and *fekr kærðæn* ‘think’ (lit. ‘idea make’). All nine nuclear verbs of European languages appear among the 20 most frequent verbs as can be observed in Table 22. In addition, BE is the most frequent verb but HAVE does not have rank 2 but rank 4, because *ʃodæn* ‘become’ (rank 2) and *kærðæn* ‘make’ (rank 3) are so frequently exploited to form compound verbs. Note that HIT (*zæðæn*, rank 18) also appears among the 20 most frequent verbs.

Concrete actions				Verbal communication
<i>Motion</i>	<i>Possession</i>	<i>Production</i>	<i>Physical contact</i>	
11 <i>ræftæn</i> ‘go’	4 <i>daftæn</i> ‘have’	3 <i>kærðæn</i> ‘make’	18 <i>zæðæn</i> ‘hit’	5 <i>goftæn</i> ‘say’
13 <i>amæðæn</i> ‘come’	10 <i>dadæn</i> ‘give’			
12 <i>residæn</i> ‘arrive’	8 <i>gereftæn</i> ‘take’			
20 <i>gozæftæn</i> ‘put, leave’	16 <i>aværðæn</i> ‘bring’			
Mental			Grammatical	
<i>Perception</i>	<i>Cognition</i>	<i>Desire</i>	<i>Modal</i>	<i>Dynamic</i>
15 <i>didæn</i> ‘see’	14 <i>danestæn</i> ‘know’	6 <i>xastæn</i> ‘want’	7 <i>tævanestæn</i> ‘can’	1 <i>budæn</i> ‘be’
17 <i>nemudæn</i> ‘show, do’			9 <i>bajestæn</i> ‘must’	2 <i>ʃodæn</i> ‘become’
18 <i>jaftæn</i> ‘find’				

Table 22. The most frequent verbs in Persian (based on Miller & Aghajanian-Stewart 2017).

The ranking is based on the raw frequencies.

Family (2014: 25) in her monograph on Persian light verbs lists 14 such verbs in descending order of frequency: *kærdæn* to do, *zædæn* to hit, *dadæn* to give, *gereftæn* to get, *daftæn* to have, *amædæn* to come, *aværdæn* to bring, *xordæn* to eat, *kešidæn* to pull, *šodæn* to become, *bordæn* to take, *ræftæn* to go, *ændaxtæn* to throw and *oftadæn* to fall. Among these, there are 5 nuclear verbs: MAKE ('do'), GO, COME, GIVE and TAKE. In addition, the two verbs HIT and EAT that have extensive patterns of polysemy in many non-European languages are used as light verbs. The verb *xordæn* 'eat' has a high frequency (rank 28, 476 pm, LN 6.2).

Turkish (nucl1301; Turkic, Common Turkic) is particularly interesting since it represents a highly agglutinating language. The display of the 20 most frequent verbs in Table 23 is based on the part of speech index in Aksan et al. (2017: 307).

Concrete actions				Verbal communication
<i>Motion</i>	<i>Possession</i>	<i>Production</i>	<i>Ingestion</i>	
6 <i>gel</i> 'come'	4 <i>al</i> 'take, get'	2 <i>et</i> 'do, make'	20 <i>ye</i> 'eat'	5 <i>de</i> 'say'
9 <i>çık</i> 'go out'	7 <i>ver</i> 'give'	3 <i>yap</i> 'do, make'		18 <i>söyle</i> 'say, tell'
11 <i>git</i> 'go'				
14 <i>geç</i> 'pass'				
Mental			Grammatical	
<i>Perception</i>	<i>Cognition</i>	<i>Desire</i>	<i>Modal</i>	<i>Dynamic</i>
8 <i>gör</i> 'see'	15 <i>bil</i> 'know'	13 <i>iste</i> 'want'		1 <i>ol</i> 'be, become'
10 <i>bul</i> 'find'	16 <i>anla</i> 'understand'			
19 <i>bak</i> 'look at'				
Other: 12 <i>çalış</i> 'work'; 17 <i>kal</i> 'stay, remain'				

Table 23. The 20 most frequent verbs in Turkish (based on Aksan et al. 2017).

Table 23 shows the stems of the verbs. (Turkish dictionaries traditionally use the infinitive marked with the suffix *-mek/-mak* as a citation form.) Because of that, unresolved ambiguities arise making some lemmas homographic. Among the verbs in Table 23, this only affects *et*, which is represented as “*et* v, n (v) to do, to make (n)

meat; flesh” in the frequency index (Aksan et al. 2017: 8). The nuclear verbs appear among the 20 most frequent verbs also in Göz (2003), which is based on a corpus with around one million words and shows the verbs lemmatized in the infinitive form. In this list, *etmek* does not appear among the 20 most frequent verbs.

All the nuclear verbs have clear parallels except that there are two verbs glossed ‘do, make’. It is worth noticing that Turkish belongs to the languages where EAT appears among the most frequent verbs. The most frequent grammatical verbs in European languages do not have direct correspondences. BE is special (see Lees 1972).

HAVE and the Modals are missing from the list of the most frequent verbs. HAVE is expressed with the existential predicate *var* ‘exist’ and its negative counterpart *yok*. Modals are described in Johansson (2009). Possibility is expressed with the suffix *-(y)A-bil* derived from the verb *bilmek* ‘to know’ preceded by the converb (CVB) *-(y)A-: gel-e-bil-* (come-CVB-know) ‘to be able to come’. This suffix can be interpreted as a morphological correspondence to CAN. Necessity presents a more complex picture and can be expressed with the two adjectives *gerik* or *lazım* (the latter of Arabic origin).

11. The 100 most frequent verbs in European languages

This section will briefly discuss what meanings are characteristic of verbs in European languages with a rank between 1 and 100. The discussion has accounted systematically for all the nuclear verbs that tend to appear among the 20 most frequent verbs. In addition, several verbs with a rank between 21 and 100 have been commented on when they belong to the same field as one of the nuclear verbs. As a background to the following discussion, the 100 most frequent verbs in English are shown in Table 24. The frequency ranking is based on Kilgariff (1998).

Grammatical verbs					
<i>Dynamic system</i>			<i>Phasal</i>	<i>Modal</i>	<i>Attempt/Success</i>
State	Change	Cause			
1 <i>be</i>	26 <i>become</i>	49 <i>help</i>	43 <i>begin</i>	2 <i>will</i>	34 <i>try</i>
	61 <i>happen</i>	67 <i>let</i>	46 <i>start</i>	5 <i>can</i>	
	74 <i>change</i>		66 <i>continue</i>	12 <i>shall</i>	
			85 <i>stop</i>	14 <i>may</i>	
				23 <i>must</i>	
				30 <i>need</i>	

(continued)

Concrete actions				
<i>Posture</i>	<i>Motion</i>		<i>Possession</i>	<i>Creation</i>
	Subject	Object		
59 <i>stand</i>	7 <i>go</i>	24 <i>put</i>	2 <i>have</i>	9 <i>make</i>
64 <i>sit</i>	16 <i>come</i>	41 <i>turn</i>	8 <i>get</i>	68 <i>produce</i>
100 <i>lie</i>	28 <i>leave</i>	42 <i>bring</i>	13 <i>take</i>	93 <i>build</i>
	40 <i>follow</i>	48 <i>set</i>	17 <i>give</i>	
	47 <i>run</i>	58 <i>lead</i>	37 <i>keep</i>	
	51 <i>move</i>	62 <i>carry</i>	38 <i>provide</i>	
	81 <i>fall</i>	86 <i>send</i>	52 <i>pay</i>	
	92 <i>return</i>		76 <i>lose</i>	
	99 <i>reach</i>		84 <i>buy</i>	
			91 <i>receive</i>	
			94 <i>spend</i>	
Mental verbs				
<i>Metalinguistic</i>	<i>Verbal Communication</i>	<i>Perception</i>	<i>Cognition</i>	<i>Desire</i>
27 <i>mean</i>	6 <i>say</i>	10 <i>see</i>	11 <i>know</i>	21 <i>want</i>
	22 <i>tell</i>	18 <i>look</i>	15 <i>think</i>	
	33 <i>ask</i>	20 <i>find</i>	56 <i>believe</i>	
	36 <i>call</i>	31 <i>feel</i>	72 <i>consider</i>	
	45 <i>write</i>	34 <i>show</i>	78 <i>expect</i>	
	63 <i>talk</i>	53 <i>hear</i>	79 <i>remember</i>	
	70 <i>require</i>	65 <i>appear</i>	87 <i>decide</i>	
	71 <i>suggest</i>	73 <i>read</i>	89 <i>understand</i>	
	75 <i>offer</i>		98 <i>learn</i>	
	82 <i>speak</i>			
	95 <i>describe</i>			
	96 <i>agree</i>			
Other fields: 4 <i>do</i> , 19 <i>use</i> , 25 <i>work</i> , 39 <i>hold</i> , 50 <i>play</i> , 54 <i>meet</i> , 55 <i>include</i> , 57 <i>allow</i> , 60 <i>live</i> , 69 <i>involve</i> , 77 <i>add</i> , 80 <i>remain</i> , 83 <i>open</i> , 88 <i>win</i> , 90 <i>develop</i> , 97 <i>increase</i> .				

Table 24. The 100 most frequent verbs in English classified into semantic fields based on Kilgarriff (1998).

11.1. Grammatical verbs

This paper has focused on verbs with a lexical meaning except BE and the modals CAN and MUST. Verbs referring to “Concrete actions” and “Mental verbs” have already been discussed in Sections 5-9. For that reason, the verbs referred to broadly as “grammatical verbs” will be commented on separately in this section, see Table 25.

In this table and Table 27, the number of languages in which a verb with the indicated meaning can be found among the 100 most frequent verbs is given within parentheses. Meanings which usually appear among the 20 most frequent verbs are shown in bold face. Lexical verbs with frequent grammaticalized meanings do not appear in the table, since information about the frequency of various uses of these verbs is available only to a limited extent (see discussion below).

Grammatical verbs					
<i>Dynamic system</i>			<i>Phasal</i>	<i>Modal</i>	<i>Attempt/Success</i>
State	Change	Cause			
BE (13)	BECOME (9)	LET (9)	BEGIN (13)	CAN (13)	TRY (11)
	CHANGE (10)	HELP (9)	CONTINUE (9)	MUST (11)	SUCCEED (6)
	HAPPEN (13)		END (10)		

Table 25. Verbs with a grammatical meaning among the 100 most frequent verbs.

The group labelled “Dynamic system” refer to State, Change (or Inchoative) and Cause. Table 26 shows how BECOME, which appears among the 100 most frequent verbs in 9 of the languages, can be related to BE (dealt with in Section 4) and to MAKE (Section 5).

Dynamic	Verb	Property	Class inclusion
State	<i>be</i>	Eve is happy	Alphie is a big star.
Change	<i>become</i>	Eve became happy	Alphie became a big star.
Cause	<i>make</i>	Bob made Eve happy	The famous director made Alphie a big star

Table 26. A sketch of the dynamic system.

BE is realized as a verb in all European languages but BECOME is realized by several different verbs with other basic meanings in some languages, for example Spanish (*ponerse* ‘put + REFL’, *hacerse* ‘make + REFL’, *volverse* ‘turn + REFL’). In Finnish, the major realization of BECOME is *tulla* ‘come’, but several other options exist. Tentatively HAPPEN is also regarded as part of the dynamic system together with CHANGE.

As mentioned, the grammatical meanings are in many cases expressed with extended uses of lexical verbs. The examples of “Cause” in Table 26 correspond to Causative-Adj and Transformation in Table 10, which illustrates uses of MAKE. A verb with this core meaning is used as a periphrastic causative in several languages. The reason why this verb does not appear in Table 25 is that only grammatical verbs

without a basic lexical meaning are shown. This is a consequence of the decision to assign only one basic meaning to each verb. The principle to stick to one meaning for each verb is done to get a manageable overview of lexical structure. The grammatical meanings of lexical verbs are accounted for both in the description of the pattern of polysemy of each verb and in the description of the systems of grammatical meaning briefly discussed here. It is an open question whether some verbs which have both a frequent lexical meaning and a frequent grammatical meaning should be treated as two separate verbs. At present, only information about the overall frequency of such verbs is available for many languages.

The function Causative-VP in Table 10 as in *He made me laugh*, which is usually referred to as the periphrastic or analytic causative, is expressed with several different verbs across languages. A variety of finer semantic distinctions can be expressed in this way but there is a major contrast between make-type and let-type causation according to Levshina (2015), who has analyzed such structures in 18 European languages. The former type is expressed by MAKE in English, the Romance languages and Greek but with GET in Finnish and Swedish. Observe that *doen*, not *maken*, can be used to express this type of causative in Dutch (see Table 11). A verb glossed ‘let’ and in some languages ‘let, leave’, which is symbolized LET, appears in 9 languages, where it can be used as a verb of causation. Tentatively, this meaning is regarded as the basic one, but LEAVE might be more appropriate for some languages. In Swedish, there is a language-specific combination of the meanings LET and SOUND in the verb *låta* (Viberg 2009). A verb with the basic meaning GIVE fulfils that function in Russian, Finnish and Greek. A verb meaning ALLOW appears in as many as 11 of the languages, but it is left open for the time being whether the basic meaning is causative or belongs the field of Verbal communication verbs (‘say that something is permitted’). According to Levshina (2015: 514), “Letting” is the prototype of European causatives, whereas make-causatives are less salient. One reason for this is that they compete with lexical causatives and in some languages also with morphological causatives as in Finnish and Hungarian. HELP represents a special type of causative meaning and a verb with that meaning is found among Top 100 in 9 languages.

Phasal verbs represent another field of grammatical verbs that tend to be frequent in European languages. All languages have at least one verb representing each of the meanings BEGIN and STOP (or FINISH, END) among the Top 100 and CONTINUE is found in this range in most of the languages. The most basic *Modals* are described in Section 4. English has unusually many very frequent modal verbs but that is rather a language-specific feature. *Attempt/Success* refers to the concepts Attempt, Success and Failure. TRY is the most frequent verb of this type and appears among the top 100 in 11 languages. Finnish is particularly rich in such verbs (Flint 1980).

The limits set up when looking at the 20 or 100 most frequent verbs are heuristic but in principle arbitrary. Based on the first limit, it could be established that the nuclear verbs could be found within this limit with only a restricted number of exceptions and that most of the exceptions had a rank only slightly above 20. When looking at the 100 most frequent verbs, no systematic attempt has been made to find out whether a certain verb could be found outside this range, which means that verbs that deviate only slightly are missed. Further studies would have to include more verbs, say with a rank up to 200, to find the most informative cut-off point.

11.2. Lexical verbs

Table 27 sums up the set of lexical verbs discussed in Sections 3-9 together with a few more language-specific verbs that will be commented on below.

Concrete actions					
Posture	Motion		Possession	Creation	Life cycle
	Subject	Object			
STAND (8)	GO (13)	LEAD (11)	HAVE (8)	MAKE (13)	LIVE (13)
SIT (7)	COME (13)	CARRY (10)	TAKE (13)	CREATE (11)	
LIE (5)	LEAVE (11)		GIVE (13)		
	FOLLOW (10)		BRING (9)		
			PAY (9)		
			SELL (7)		
			GET (9)		
			RECEIVE (9)		
Mental verbs					
Metalinguistic	Verbal communication		Perception	Cognition	Desire
MEAN (7)	SAY (13)		SEE (13)	KNOW (13)	WANT (13)
	SPEAK (12)		HEAR (12)	THINK (12)	
	WRITE (13)		SHOW (12)	BELIEVE (12)	
	ASK (13)		FEEL (12)	DECIDE (11)	
	ANSWER (7)		SEEM (13)	UNDERSTAND (10)	
	TELL/narr. (8)				
Other verbs: USE (12), WORK (12)					

Table 27. Lexical verbs that appear among the 100 most frequent verbs

The field “Life cycle” has been added, which contains LIVE in the sense ‘be alive’ together with verbs such as GROW and DIE. Among “Other verbs”, USE might have been considered as a grammatical verb categorized as Instrumental, because of the alternation *Peter used a crowbar to open the window* / *Peter opened the window with a crow bar*. A verb glossed WORK appears among the 100 most frequent verbs in 12 languages. The sense ‘do a job’ appears to be basic but that requires further study.

11.3. Language-specific and variable features

The lists of the 100 most frequent verbs can also be used to identify language-specific features in European languages. Verbs that have a high frequency only in one language or in a restricted group of languages are language-specific. One example of that is DO discussed in Section 5. English *tell* discussed in Section 8 should also be regarded as language-specific. Another type of language-specific characteristic is the lack of a verb that is frequent in most of the other European languages such as the lack of a frequent verb meaning BRING in Swedish and Welsh as described in Section 7. Some features are variable without being language-specific since they are characteristic of certain groups of languages. An example is *Postural verbs*, which have a high frequency in Germanic and Slavic languages, in which all the three verbs SIT, STAND and LIE appear among the 100 most frequent verbs, whereas other European languages only occasionally have a postural verb in this frequency range. One important reason for this is the frequent use of the Postural verbs to indicate the location of physical objects in the Slavic and Germanic languages, except in English where location most frequently is indicated with BE and *sit*, *stand* and *lie* have a lower frequency than their correspondences in Swedish and German, where postural verbs are among the 50 most frequent verbs (see Viberg 2013a). Among verbs of Motion, ENTER and EXIT, which are characteristic of V-framed languages, have already been discussed in Section 6.4. Among the verbs of Possession, verbs meaning GET (in the sense ‘receive’) have a particularly high frequency in Swedish (*få*), Finnish (*saada*) and Estonian (*saama*), where these verbs have rank 5 or lower, see the articles in Lenz & Rawoens (2012). The verbs glossed RECEIVE have a much lower frequency even if they appear among the 100 most frequent verbs. Language-specific and other variable features that are characteristic of the 100 most frequent verbs are displayed in Table 28.

Semantic field	Variable features
Production	DO vs. MAKE
Spatial	
<i>Location</i>	BE vs. Postural location verbs
<i>Motion</i>	Assisted vs. Non-assisted motion The deictic component of COME A general verb PUT vs. other alternatives Path verbs
Possession	HAVE vs. BE A frequent, polysemous GET TAKE vs. TAKE ^{CM} TAKE vs. BRING GIVE → LET
Verbal	SAY vs. TELL ASK = REQUEST vs. two separate verbs
Perception	
Experiences	Separate verbs in all sense modalities vs. FEEL extended to various extents
Cognition	The extension of KNOW Contrasts between verbs of thinking

Table 28. Some important variable features in European languages.

12. Why are nuclear and other basic verbs important?

Viberg (1993) claimed that nuclear verbs tend to have many hyponyms, to have characteristic and elaborate patterns of polysemy and to have a strong tendency to develop grammatical meanings. Grammaticalization is an aspect of ordinary language change. Nuclear verbs also play a central role in lexical progression and regression, i.e. in developing and in reduced lexical systems. My own interest in research on basic vocabulary arose when I studied the acquisition of Swedish as a second language by children in preschool and early primary school. It turned out that frequency of occurrence was an important factor for the acquisition of verbs. At early stages, several of the nuclear verbs were relied on heavily in oral production. Such verbs were overused in terms of frequency in comparison to L1 children of the same age and had a strong tendency to be semantically overextended (Viberg 1993, 1999b). Together

the nuclear verbs cover major distinctions in the semantic space of actions in a systematic way. Learners can make themselves understood to a certain degree even if the reference is less precise than when more specific verbs are used by native speakers. As a set, nuclear verbs cover most of the different types of argument structures characteristic of an individual language and in several respects serve as syntactic prototypes (Viberg 1993, 2002).

More recently, several studies have appeared that point to the importance of (what I refer to as) nuclear verbs in first language acquisition. Maouene et al. (2008, 2016) related the early development of the verb lexicon to embodiment and looked at the relationship between early verbs, body parts and syntactic frames (argument structure) in corpora of naturalistic speech by two groups of children 20 and 28 months old and of their mothers' speech in interactions with their children. With growing age, there was a clear development from relatively many mouth verbs (e.g., *bite*, *blow*, *drink*), to more hand verbs (e.g., *put*, *take*, *write*), to verbs less closely associated with the body. What is especially relevant for the study of basic verbs is the relationships that were found between syntactic frames and the most frequent verbs appearing in a specific frame as shown in Table 29.

28-month-olds and mothers (both groups)			20-month-olds
<i>Syntactic frame</i>	<i>Body region</i>	<i>Most frequent verb(s)</i>	<i>Body region</i>
V	HEAD	<i>go, look</i>	ARM & HEAD
V Locative	LEG	<i>go</i>	LEG
V NP	ARM	<i>get, eat</i>	HEAD & ARM
V NP Locative	ARM	<i>put</i>	ARM
V NP NP	ARM	<i>give</i>	---
V Sentence	HEAD	<i>say, think</i>	---

Table 29. The relationships between syntactic frames and their most frequent verbs in child language (freely based on Maouene et al., 2016).

The result was similar for 28-month-olds and mothers of both age groups, whereas 20-month-olds lacked examples of the two most complex syntactic frames. The general conclusion was that syntactic development was driven both by high frequency and embodiment, which was shown to play an independent role for the type of less frequent verbs that appeared in the frames. Three nuclear verbs are found among the

most frequent verbs in Table 29; *go*, *give*, *say* and perhaps also *think* are nuclear verbs. High frequency plays a role also for some typologically variable verbs such as *put*, which is regarded as a language-specific basic verb. What is specific about nuclear verbs is that they have a high frequency in all European languages and tentatively also in many other languages.

Davis & Landau (2021) present a detailed study how perception and mental (my “cognitive”) verbs interact with sentence frames in children’s language development. There is a special focus on embedded clauses (cf. V + Sentence, above), since such clauses are key to referencing mental states for both verb classes. The study is based on a large sample of child-produced and parent-produced utterances (age range 2;0-4;11). The most frequent verbs are shown in Table 30.

Perception	N	Mental	N	Other
<i>see</i>	2175	<i>know</i>	1216	296
<i>look</i>	1190	<i>think</i>	386	
<i>watch</i>	297	<i>forget</i>	84	
<i>hear</i>	170	<i>guess</i>	70	
TOTAL			5884	

Table 30. The most frequent child-produced perception and mental verbs
(based on Davis & Landau 2021: 35).

The nuclear verbs *see* and *know* dominate in their respective fields, even though some frequent language-specific verbs such as *watch* are also used from early on, but not as frequently. Contrary to what has often been expected, mental verbs turned out to lead the development of embedded clauses, because the children produced embedded clauses for mental verbs more often, when they started to use such clauses more regularly around age three. Perception verbs at this age were frequently used with less complex sentence frames and less abstract meaning.

13. Brief summary and Conclusion

This article can partly be regarded as a replication of an earlier study (Viberg 1993) based on corpora of a restricted size (around 1 million words). Further support has been provided for the identification of nine nuclear verbs and a couple of grammatical verbs as the most basic lexical verbs in European languages, see Table 31.

a. Reflecting universal tendencies: The nuclear verbs						
<i>Motion</i>	<i>Possession</i>	<i>Production</i>	<i>Verbal com.</i>	<i>Perception</i>	<i>Cognition</i>	<i>Desire</i>
GO	GIVE	MAKE	SAY	SEE	KNOW	WANT
COME	TAKE					
b. Dynamic system		c. Areally specific basic verbs		d. Subareally specific basic verb		
<i>Copula</i>		<i>Modal verbs</i>		<i>Possession: state</i>		
BE		CAN		HAVE		
		MUST				

Table 31. Basic verbs in European languages.

In most of the 13 tested European languages, the verbs in Table 31 are among the 20 most frequent verbs and have a frequency above 1100 pm (LN 7.0). The only clear exception is HAVE, which is completely lacking in some languages. There are also a few languages where MUST is not expressed by a verb. It has only been possible to sketch the analysis of individual semantic fields the basic verbs belong to, but more extended discussion can be found in the references to earlier contrastively oriented work by the author and to work by other scholars in Sections 3-9. The next step, which has only been sketched, is to investigate to what extent there are parallels in non-European languages.

The extension of the study to the 100 most frequent verbs points to possible further generalizations. The approach also makes it possible to identify frequent but language-specific basic verbs. Such verbs make up a significant part of the typological lexical profile of an individual language (Viberg 2006a) and are interesting among other things because they tend to represent serious problems for second language learners.

In non-European languages, the concepts associated with the nuclear verbs and other basic verbs may be realized as frequent and basic expressions that are not verbs or are not simple verbs but multiwords or morphologically complex verbs. From a theoretical point of view, it would be interesting to base the comparison on word stems or on morphemes (cf. Section 10 on Turkish). For polysynthetic languages, it appears to be necessary to look at morphemes. Compound words and lexicalized phrases represent another problem but do not in general appear among the 100 most frequent verbs in European languages. One example of a different kind of language is Persian, which was briefly discussed for comparison in Section 10. Several of the characteristic compound verbs are treated as separate entries, for example *fekr kærðæn* ‘think’ and *pejda kærðæn* ‘find’ (*kærðæn* ‘do/make’). This is obviously motivated by

the fact that there are very few simple verbs in Persian. This could motivate to base the comparison directly on concepts such as THINK or FIND. Partly, this has already been done in the present study, but I must leave it to further studies to refine such theoretical issues. (I made an early attempt by looking at the way the sense modality hierarchy of perception verbs is reflected in morphological complexity in Viberg 1983b.) There is no denying that cultural differences lead to certain concepts being more established in a way that is reflected in the distinctions expressed by words belonging to a certain semantic field, but a hypothesis worth testing is that certain concepts are so basic that they tend to be expressed by very frequent expressions in most or all languages. Since European languages have a large number of simple verbs, it is natural that basic events are expressed by frequent simple verbs, but the analysis must be adjusted to accommodate languages where compound verbs are the norm (except for a restricted number of simple verbs) or languages with different types of morphological structure than European languages such as polysynthetic languages.

Acknowledgements

I wish to thank the two anonymous reviewers and Francesca Di Garbo for insightful comments and Nick Ellis, Jonė Grigaliūnienė and Maria Koptjevskaja-Tamm for help with various aspects of the analysis of Welsh, Lithuanian and Russian, respectively. Remaining shortcomings are my own responsibility.

Abbreviations

1 = 1 st person	CM = complement marker	PROG = progressive
2 = 2 nd person	DEF = definite	PRS = present
3 = 3 rd person	FUT = future	PTCP = participle
ACC = accusative	PASS = passive	SG = singular
AD = adessive	PRF = perfect	VN = verbal noun

References

- Aijmer, Karin. 2015. Pragmatic markers. In Karin Aijmer & Christoph Rühlemann (eds.), *Corpus pragmatics. A handbook*, 195–218. Cambridge: Cambridge University Press.

- Aijmer, Karin. 2019. What is happening to *must* in Present-day English? A contrastive perspective on a declining modal auxiliary. In Harumi Sawada, Yoshio Nitta & Masa-aki Yamanashi (eds.), *Bamen to shutaisei/shukansei* [Situation and subjectivity], 387–402. Tokyo: Hituzi Syobo.
- Aksan, Yeşim, Mustafa Aksan, Ümit Mersinli & Umut Ufuk Demirhan. 2017. *A frequency dictionary of Turkish*. London: Routledge.
- Altenberg, Bengt & Karin Aijmer. 2000. The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In Christian Mair & Marianne Hundt (eds.), *Corpus linguistics and linguistic theory*, 15–23. Amsterdam: Rodopi.
- Ballmer, Thomas & Waltraud Brennenstuhl. 1981. *Speech act classification. A study in the lexical analysis of English speech activity verbs*. Berlin: Springer-Verlag.
- Bertinetto, Pier Marco, Luca Ciucci & Denis Creissels. 2026a. *Non-verbal predication in the world's languages: A typological survey. Volume 1: Eurasia, North America, South America*. Berlin: De Gruyter Mouton.
- Bertinetto, Pier Marco, Luca Ciucci & Denis Creissels. 2026b. *Non-verbal predication in the world's languages: A typological survey. Volume 2: Africa, Austronesia, Papunesia, Australia*. Berlin: De Gruyter Mouton.
- Bouveret, Myriam (ed.). 2021. *Give constructions across languages* [Constructional Approaches to Language 29]. Amsterdam: Benjamins. <https://doi.org/10.1075/cal.29>.
- Buck, Carl Darling. 1949. *A dictionary of selected synonyms in the principal Indo-European languages*. Chicago & London: The University of Chicago Press.
- Camacho, José. 2012. *Ser* and *estar*: The individual/stage-level distinction and aspectual predication. In José Ignacio Hualde, Antxon Olarrea & Erin O'Rourke (eds.), *The Handbook of Hispanic Linguistics*, 453–475. Oxford: John Wiley & Sons.
- Čermáková, Anna, Jarle Ebeling & Signe Oksefjell Ebeling. 2023. 'Be' verbs in a contrastive perspective: The case of BÝT, BE and VÆRE. *Nordic Journal of English Studies* 22(2). 27–58.
- Collinder, Björn. 1955. *Fenno-Ugric vocabulary*. Stockholm: Almqvist & Wiksell.
- Davies, John. 1981. *Kobon* [Lingua Descriptive Studies]. Amsterdam: North-Holland.
- Davis, E. Emory & Barbara Landau. 2021. Seeing and Believing: The relationship between perception and mental verbs in acquisition. *Language Learning and Development* 17(1). 26–47. <https://doi.org/10.1080/15475441.2020.1862660>.
- DeLancey, Scott. 1985. The analysis-synthesis-lexis cycle in Tibeto-Burman: A case study in motivated change. In John Haiman (ed.), *Iconicity in syntax*, 367–89. Amsterdam: Benjamins.

- Divjak, Dagmar. 2019. *Frequency in language: Memory, attention and learning*. Cambridge: Cambridge University Press.
- Dixon, R. M. W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.
- Ebeling, Signe Oksefjell. 2017. Bringing home the bacon! A contrastive study of the cognates *bring/bringe* in English and Norwegian. *Kalbotyra* 70. 104–126. <https://doi.org/10.15388/Klbt.2017.11193>.
- Egan, Thomas. 2021. TELLING in English, Norwegian and French: A three-way contrast. *BLS* 11(1). 47–68.
- Evans, Nicholas & David Wilkins. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language* 76(3). 546–592.
- Family, Neiloufar. 2014. *Semantic spaces of Persian light verbs. A constructionist account* [Brill's Studies in South and Southwest Asian Languages Series 6]. Leiden: Brill.
- Ferguson, Charles. 1976. The Ethiopian language area. In M. Lionel Bender, J. Donald Bowen, Robert L. Cooper & Charles A. Ferguson (eds.), *Language in Ethiopia*, 63–76. Oxford: Oxford University Press.
- Fetzer, Anita & Marjut Johansson. 2012. Cognitive verbs in context. A contrastive analysis of English and French argumentative discourse. In Stefania Marzo, Kris Heylen & Gert de Sutter (eds.), *Corpus studies in contrastive linguistics*, 89–115. Amsterdam: Benjamins. <https://doi.org/10.1075/bct.43.05fet>.
- Fillmore, Charles J. 1997. *Lectures on deixis* [CSLI lecture notes 65]. Stanford: CSLI Publications.
- Flint, Aili. 1980. *Semantic structure in the Finnish lexicon: verbs of possibility and sufficiency* [Suomalaisen kirjallisuuden seuran toimituksia 0355–1768; 360]. Helsinki: Suomalaisen Kirjallisuuden Seura.
- Foley, William A. 1986. *The Papuan languages of New Guinea*. Cambridge: Cambridge University Press.
- Fortescue, Michael. 2001. Thoughts about thought. *Cognitive Linguistics* 12(1). 15–45.
- Frommherz, Yannick. 2022. Thinking things in German versus Swedish. A cross-linguistic comparison of verbs of thinking in two genetically close languages. *Studia Linguistica* 76(2). 464–506.
- Gilquin, Gaëtanelle & Åke Viberg. 2009. How to make do with one verb: a corpus-based contrastive study of DO and MAKE. *Revue française de linguistique appliquée* 14(1). 67–82.
- Goddard, Cliff. 2001. Lexico-semantic universals: A critical overview. *Linguistic Typology* 5. 1–65.

- Goddard, Cliff. 2003. Thinking across languages and cultures: six dimensions of variation. *Cognitive Linguistics* 14(2/3), 109–140 Special issue edited by Gary B. Palmer, Cliff Goddard & Penny Lee.
- Göz, İlyas. 2003. *Yazılı türkçenin kelime sıklığı sözlüğü* [A Frequency Dictionary of Written Turkish]. Ankara: Türk Dil Kurumu.
- Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.
- Hammarberg, Björn & Åke Viberg. 1977. The place-holder constraint, language typology, and the teaching of Swedish to immigrants. *Studia Linguistica* 31. 106–163.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2026. Glottolog 5.3. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: <http://glottolog.org> (Accessed 2026-07-06).
- Hansen, Björn & Ferdinand de Haan (eds.). 2009. *Modals in the languages of Europe*. Berlin: Mouton de Gruyter.
- Harkins, Jean. 1995. Desire in language and thought: A study in cross-cultural semantics. Canberra: Australian National University. (Doctoral Dissertation.) <https://doi.org/10.25911/5d7638e7c1b34>.
- Harras, Gisela, Edeltraud Winkler, Sabine Erb & Kristel Proost. 2004. *Handbuch deutscher Kommunikationsverben 1. Wörterbuch*. Berlin: de Gruyter.
- Harras, Gisela, Kristel Proost & Edeltraud Winkler. 2007. *Handbuch deutscher Kommunikationsverben 2. Lexikalische Strukturen*. Berlin: de Gruyter.
- Haspelmath, Martin. 2013. ‘Want’ complement subjects. In Matthew Dryer & Martin Haspelmath (eds.), *WALS Online*. Available online at: <http://wals.info/chapter/124> (Accessed 2026-07-06)
- Heine, Bernd. 1997. *Possession. Cognitive sources, forces, and grammaticalization*. Cambridge: Cambridge University Press.
- Heine, Bernd & Tania Kuteva. 2002. *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.
- Hoddinott, William G. & Frances M. Kofod. 1976. Djamindjungan. In R. M. W. Dixon (ed.), *Grammatical categories in Australian languages*, 698–704. New Jersey: Humanities Press.
- Isačenko, Alexander V. 1974. On ‘have’ and ‘be’ languages. In Michael S. Flier (ed.), *Slavic forum: Essays in linguistics and literature*, 43–77. The Hague: Mouton.
- Jakobson, Roman. 1971. *Selected Writings*, vol. 2. The Hague: Mouton.

- Johansson, Lars. 2009. Modals in Turkic. In Björn Hansen & Ferdinand de Haan (eds.), *Modals in the languages of Europe*, 487–510. Berlin: Mouton de Gruyter.
- Juvonen, Päivi. 2016. Making do with minimal lexica. Light verb constructions with MAKE/DO in pidgin lexica. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*, 223–248. Berlin: de Gruyter Mouton.
- Kilgariff, Adam. 1998. BNC lists. <https://www.kilgariff.co.uk/bnc-readme.html> (Accessed 2026-07-06).
- King, Gareth. 2016. *Modern Welsh. A comprehensive grammar*. London: Routledge.
- Kopecka, Anetta. 2006. The semantic structure of motion verbs in French: Typological perspectives. In Maya Hickman & Stéphanie Robert (eds.), *Space in languages. Linguistic systems and cognitive categories*, 83–101. Amsterdam: Benjamins.
- Kopecka, Anetta & Bhuvana Narasimhan (eds.). 2012. *Events of Putting and Taking*. Amsterdam: Benjamins.
- Kuiper, Alberta & William R. Merrifield. 1975. Diuxi Mixtec verbs of motion and arrival. *International Journal of American Linguistics* 41(1). 32–45.
- Langlotz, Andreas. 2006. *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. Amsterdam: Benjamins.
- Lees, Robert B. 1972. The Turkish copula. In John W. M. Verhaar (ed.), *The Verb 'Be' and its synonyms. Philosophical and grammatical studies*, Part 5, 64–73. Dordrecht: Reidel Publishing Company.
- Lenz, Alexandra, Fabian Fleißner, Agnes Kim & Stefan Michael Newerkla. 2020. GIVE as a German PUT verb – Analyses on Slavic-German language contact. *Journal of Linguistic Geography* 8(2). 67–81.
- Lenz, Alexandra N. & Gudrun Rawoens (eds.). 2012. The art of getting: GET verbs in European languages from a synchronic and diachronic point of view. Special issue of *Linguistics* 50(6).
- Leonetti, Manuel, Silvia Gumiel-Molina & Isabel Pérez-Jiménez. 2015. *Ser and estar: Outstanding questions*. In Isabel Pérez-Jiménez, Manuel Leonetti & Silvia Gumiel-Molina (eds.), *New perspectives on the study of ser and estar*, 1–20. Amsterdam: Benjamins.
- Levin, Beth. 1993. *English verb classes and alternations*. Chicago: The University of Chicago Press.
- Levshina, Natalia. 2015. European analytical causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487–520. <https://doi.org/10.1515/flin-2015-0017>.

- Lewandowski, Wojciech. 2014. Deictic verbs: typology, thinking for speaking and SLA. *SKY Journal of Linguistics* 27. 43–65.
- Liljegren, Henrik. 2010. Where have all the verbs gone? On verb stretching and semi-words in Indo-Aryan Palula. *Himalayan Linguistics* 9(1). 50–79.
- Lonsdale, Deryle & Yvon Le Bras. 2009. *A frequency dictionary of French*. London: Routledge.
- Lord, Carol. 1993. *Historical change in serial verb constructions* [Typological Studies in Language 26]. Amsterdam: Benjamins. <https://doi.org/10.1075/tsl.26>.
- Loveday, Leo. 1986. *Explorations in Japanese Sociolinguistics*. Amsterdam: Benjamins.
- Margetts, Anna, Katharina Haude, Nikolaus P. Himmelmann, Dagmar Jung, Sonja Riesberg, Stefan Schnell, Frank Seifart, Harriet Sheppard & Claudia Wegener. 2022. Cross-linguistic patterns in the lexicalisation of *bring* and *take*. *Studies in Language* 46(4). 934–993. <https://doi.org/10.1075/sl.19088.mar>.
- Margetts, Anna, Sonja Riesberg & Birgit Hellwig (eds.) (2022). *Caused accompanied motion: Bringing and taking events in a cross-linguistic perspective* [Typological Studies in Language 134]. Amsterdam: Benjamins.
- Maouene, Josita, Shohei Hidaka, & Linda B. Smith. 2008. Body parts and early-learned verbs. *Cognitive Science* 32. 1200–1216.
- Maouene, Josita, Nitya Sethuraman, Mounir M. Maouene & Sango Otieno. 2016. Contingencies between verbs, body parts, and argument structures in maternal and child speech: a corpus study. *Language and Cognition* 8. 237–282. <https://doi.org/10.1017/langcog.2014.48>.
- Martinková, Michaela & Markéta Janebová. 2024. English and Czech venitive Verbs in contrast: Deictic or not?. *Nordic Journal of English Studies* 23(2). 34–60.
- Matsumoto, Yo (ed.). 2025. *Motion event descriptions across languages. Volume 1: Case studies of linguistic representations of motion*. Berlin: De Gruyter Mouton.
- McGregor, William B. 2002. *Verb classification in Australian languages*. Berlin: Mouton de Gruyter.
- Meex, Birgitta. 2020. Patterns of path encoding in German. In Yo Matsumoto & Kazuhiro Kawachi (eds.), *Broader perspectives on motion event descriptions*, 63–104. Amsterdam: Benjamins.
- Miller, Corey & Karineh Aghajanian-Stewart. 2017. *A Frequency dictionary of Persian*. London: Routledge.
- Mulder, Gijs. 2017. When feeling is thinking. A lexical-semantic analysis of evidential and epistemic predicates in Spanish. In Juana I. Marín Arrese et al. (eds.),

- Evidentiality Revisited: Cognitive grammar, functional and discourse-pragmatic perspectives*, 105–121. Amsterdam: Benjamins.
- Naughton, James & Karen von Kunes. 2021. *Czech. An essential grammar*. 2nd edn. London: Routledge.
- Newman, John. 1996. *Give. A cognitive linguistic study*. Berlin: Mouton de Gruyter.
- Newman, John (ed.). 1998. *The linguistics of giving*. Amsterdam: Benjamins.
- Newman, John (ed.). 2009. *The linguistics of eating and drinking*. Amsterdam: Benjamins.
- Nolan, Brian, Gudrun Rawoens & Elke Diedrichsen (eds.). 2015. *Causation, permission, and transfer. Argument realisation in GET, TAKE, PUT, GIVE and LET verbs* [Studies in Language Companion Series, 167]. Amsterdam: Benjamins. <https://doi.org/10.1075/slcs.167>.
- Norcliffe, Elisabeth & Asifa Majid. 2024a. Verbs of perception: A quantitative typological study. *Language* 100(1). 81–123.
- Norcliffe, Elisabeth & Asifa Majid. 2024b. Word formation patterns in the perception domain: a typological study of cross-modal semantic associations. *Linguistic Typology* 28(3). 419–459.
- Nuyts, Jan. 2014. Analyses of the modal Meanings. In Jan Nuyts & Johan van der Auwera (eds.), *The Oxford Handbook of Modality and Mood*, 31–49. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199591435.001.0001>.
- Palmer, Gary B., Cliff Goddard & Penny Lee (eds.). 2003. Talking about thinking across languages. Special issue of *Cognitive Linguistics* 14(2/3).
- Pawley, Andrew. 1993. A language which defies description by ordinary means. In William Foley (ed.), *The role of theory in language description*, 87–129. Berlin: Mouton de Gruyter.
- Pawley, Andrew. 2006. Where have all the verbs gone? Remarks on the organisation of languages with closed verb classes. (Paper presented at the 11th Biennial Rice University Linguistics Symposium, 16-18 March 2006.) http://www.ruf.rice.edu/~lingsymp/Pawley_paper.pdf (Accessed 2026-07-06).
- Pawley, Andrew. 2020. The depiction of sensing events in English and Kalam. In Helen Bromhead & Zhengdao Ye (eds.), *Meaning, life and culture: In conversation with Anna Wierzbicka*. ANU Press. Canberra: Australia. <https://doi.org/10.22459/M.L.C.2020.20>.
- Raag, Raimo. 2014. *Estniska från grunden* [Finsk-ugriska läromedel FU 25]. Uppsala: Uppsala University.

- Reesink, Ger. 2013. Expressing the GIVE event in Papuan languages: A preliminary survey. *Linguistic Typology* 17(2). 217–266.
- Ricca, Davide. 1993. *I verbi deittici di movimento in Europa: una ricerca interlinguistica* [Deictic Motion Verbs in Europe: A Cross-Linguistic Study]. Florence: La Nuova Italia Editrice.
- Rice, Sally. 1998. Giving and taking in Chipewyan: the semantics of THING-marking classificatory verbs. In John Newman (ed.), *The linguistics of giving*, 97–134. Amsterdam: Benjamins.
- Rudzka-Ostyn, Brygida. 1989. Prototypes, schemas and cross-category correspondences: The case of *ask*. *Linguistics* 27. 613–661.
- Rudzka-Ostyn, Brygida. 1995. Metaphor, schema, invariance: The case of verbs of answering. In Louis Goossens, Paul Pauwels, Brygida Rudzka-Ostyn, Anne-Marie Simon-Vandenberghe & Johan Vanparys (eds.), *By Word of Mouth*, 205–243. Amsterdam: Benjamins.
- San Roque, Lila, Kobin H. Kendrick, Elisabeth Norcliffe, Penelope Brown, Rebecca Defina, Mark Dingemanse, Tyko Dirksmeyer, Nicholas J. Enfield, Simeon Floyd, Jeremy Hammond, Giovanni Rossi, Sylvia Tufvesson, Saskia van Putten & Asifa Majid. 2015. Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics* 26(1). 31–60.
- Sarda, Laure & Ludovica Lena (eds.). 2023. *Existential constructions across languages. Forms, meanings and functions*. Amsterdam: Benjamins.
- Schultze-Berndt, Eva. 2000. *Simple and complex verbs in Jamingung*. Nijmegen: MPI Series in Psycholinguistics.
- Schultze-Berndt, Eva. 2008. What do “do” verbs do? The semantic diversity of generalised action verbs. In Elisabeth Verhoeven, Stavros Skopeteas, Yong-Min Shin, Yoko Nishina & Johannes Helmbrecht (eds.), *Studies on grammaticalization*, 185–208. Berlin: De Gruyter.
- Stassen, Leon. 1997. *Intransitive predication*. Oxford: Clarendon Press.
- Stassen, Leon. 2009. *Predicative possession*. Oxford: Oxford University Press.
- Stassen, Leon. 2013a. Predicative Possession. In Matthew S. Dryer & Martin Haspelmath (eds.), *WALS Online*. Available online at: <http://wals.info/chapter/117> (Accessed 2026-07-06).
- Stassen, Leon. 2013b. Zero Copula for Predicate Nominals. In Matthew S. Dryer & Martin Haspelmath (eds.), *WALS Online*. Available online at: <http://wals.info/chapter/120> (Accessed 2026-07-06).

- Stassen, Leon. 2013c. Predicative Adjectives. In Matthew S. Dryer & Martin Haspelmath (eds.), *WALS Online*. Available online at: <http://wals.info/chapter/118> (Accessed 2026-07-06).
- Sussex, Roland & Paul Cubberley. 2006. *The Slavic languages*. Cambridge: Cambridge University Press.
- Sweetser, Eve. 1990. *From etymology to pragmatics*. Cambridge: Cambridge University Press.
- Talmy, Leonard. 1985. Lexicalization patterns: semantic structure in lexical forms. In Timothy Shopen (ed.), *Language typology and syntactic description. III. Grammatical categories and the lexicon*, 57–149. Cambridge: Cambridge University Press.
- Thornell, Christina. 2002. The uses of the Sango verb *mû* ‘transfer’. *Afrika und Übersee* 85. 267–299.
- Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. Istanbul: European Language Resources Association (ELRA).
- Tieken-Boon van Ostade, Ingrid, Marijke van der Wal & Arjan van Leuvensteijn (eds.) 1998. *DO in English, Dutch and German. History and present-day variation*. Amsterdam: Stichting Neerlandistiek & Nodus Publikationen.
- van der Auwera, Johan & Vladimir A. Plungian 1998. Modality’s semantic map. *Linguistic Typology* 2(1). 79–124.
- Van Valin, Robert D. & Randy J. LaPolla. 1997. *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press.
- Verhaar, John W. M. (ed.). 1967-1972. *The Verb ‘Be’ and its synonyms. Philosophical and grammatical studies*, Part 1-6 [Foundations of Language Supplementary Series, vols. 1, 6, 8, 9, 14 & 16.]. Dordrecht: Reidel Publishing Company.
- Viberg, Åke. 1983a. The verbs of perception: a typological study. *Linguistics* 21(1). 123–162.
- Viberg, Åke. 1983b. A universal lexicalization hierarchy for the verbs of perception. In Fred Karlsson (ed.), *Papers from the 7th Scandinavian Conference of General Linguistics*, 260–275. Helsinki: Department of General Linguistics, University of Helsinki.
- Viberg, Åke. 1993. Crosslinguistic perspectives on lexical organization and lexical progression. In Kenneth Hyldenstam & Åke Viberg (eds.), *Progression and regression in language*, 340–385. Cambridge: Cambridge University Press.

- Viberg, Åke. 1999a. The polysemous cognates Swedish *gå* and English *go*. Universal and language-specific characteristics. *Languages in Contrast* 2(1). 87–113.
- Viberg, Åke. 1999b. Lexical development in a second language. In Guus Extra & Ludo Verhoeven (eds.), *Bilingualism and migration*, 165–185. Berlin: Mouton de Gruyter.
- Viberg, Åke. 2001. The verbs of perception. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals. An international handbook*, 1294–1309. Berlin: De Gruyter.
- Viberg, Åke. 2002. Basic verbs in lexical progression and regression. In Petra Burmeister, Thorsten Piske & Andreas Rohde, A. (eds.), *An integrated view of language development. Papers in honor of Henning Wode*, 109–134. Trier: Wissenschaftlicher Verlag.
- Viberg, Åke. 2003. The polysemy of the Swedish verb *komma* ‘come’: A view from translation corpora. In Kasia M. Jaszczolt & Ken Turner (eds.), *Meaning through language contrast*, vol. 2, 75–105. Amsterdam: Benjamins.
- Viberg, Åke. 2005. The lexical typological profile of Swedish mental verbs. *Languages in Contrast* 5(1). 121–157.
- Viberg, Åke. 2006a. Towards a lexical profile of the Swedish verb lexicon. *Sprachtypologie und Universalienforschung* 59(1): 103–129. Thematic issue on: The Typological Profile of Swedish.
- Viberg, Åke. 2006b. What one verb can do: The Swedish verb *göra* in a crosslinguistic perspective. In Mickael Suominen et al. (eds.), *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday. Special Supplement to SKY Journal of Linguistics* 19. Turku, The Linguistic Association of Finland, 243–257.
- Viberg, Åke. 2009. The meaning patterns of Swedish *låta* ‘sound’; ‘let’. A corpus-based contrastive study. In Petra Bernardini, Verner Egerland & Jonas Granfeldt (eds.), *Mélanges plurilingues offerts à Suzanne Schlyter à l’occasion de son 65ème anniversaire*. [Études Romanes de Lund 85], 455–480. Centre for Languages and Literature, Lund University.
- Viberg, Åke. 2010. Basic verbs of possession. In Maarten Lemmens (ed.), *Unison in multiplicity: Cognitive and typological perspectives on grammar and lexis* [CogniTExtes 4]. <http://cognitextes.revues.org/308> (Accessed 2026-07-06).
- Viberg, Åke. 2012. Basic verbs in typological perspective. In Marleen Van Peteghem, Peter Lauwers, Els Tobbak, Annemie Demol & Laurence De Wilde (eds.), *Le verbe en verve. Réflexions sur la syntaxe et la sémantique verbales*, 255–273. Ghent: Academia Press.

- Viberg, Åke. 2013a. Posture verbs. A multilingual contrastive study. *Languages in Contrast* 13(2). 139–169.
- Viberg, Åke. 2013b. Seeing the lexical profile of Swedish through multilingual corpora. The case of Swedish *åka* and other vehicle Verbs. In Karin Aijmer & Bengt Altenberg (eds.), *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*, 25–56. Amsterdam: Benjamins.
- Viberg, Åke. 2013c. Ord om konsten att förflytta sig: Svenskans gå, åka och köra i tvärspråkligt perspektiv. In Björn Lindblom (ed.), *Text, tal & tecken. Några perspektiv inom språkforskningen* [KVHAA Konferenser 83], 116–138. Stockholm: Kungl. Vitterhets Historie och Antikvitets Akademien.
- Viberg, Åke. 2015a. Contrasts in construction and semantic composition: Crosslinguistic perspectives on the verbs of Putting in English and Swedish. In Signe Oksefjell Ebeling & Hilde Hasselgård (eds.), *Cross-Linguistic Perspectives on Verb Constructions*, 222–253. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Viberg, Åke. 2015b. Sensation, Perception and Cognition. Swedish in a typological-contrastive perspective. *Functions of Language* 22(1). 96–131.
- Viberg, Åke. 2016a. What happens in translation? A comparison of original and translated texts containing verbs meaning SIT, STAND and LIE in the English-Swedish Parallel Corpus (ESPC). *Nordic Journal of English Linguistics (NJES)* 15(3), 102–148. Available online at: <https://publicera.kb.se/njes/article/view/26824/21895> (Accessed 2026-07-06).
- Viberg, Åke. 2016b. Asking and answering: A contrastive study of English and Swedish basic communication verbs. *Linguistics Beyond and Within* 2. 180–212. Available online at: <https://czasopisma.kul.pl/index.php/LingBaW/article/view/5646> (Accessed 2026-07-06).
- Viberg, Åke. 2016c. Polysemy in action: The Swedish verb *slå* ‘hit, strike, beat’ in a crosslinguistic perspective. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*. [Cognitive Linguistics Research 58], 177–222. Berlin & Boston: de Gruyter Mouton.
- Viberg, Åke. 2017a. Contrasts in morphology. The case of UP/DOWN and IN/OUT as bound morphemes in Swedish and their English correspondents. In Markéta Janebová, Ekaterina Lapshinova-Koltunski & Michaela Martinková (eds.), *Contrasting English and other languages through corpora*, 32–74. Cambridge: Cambridge Scholars.

- Viberg, Åke. 2017b. SAYING, TALKING and TELLING. Basic verbal communication verbs in Swedish and English. In Thomas Egan & Hildegunn Dirdal (eds.), *Cross-linguistic correspondences: From lexis to genre*, 37–74. Amsterdam: Benjamins
- Viberg, Åke. 2017c. Finding a model for contrastive lexical semantics: A look at verbal communication verbs. *Linguistics Beyond and Within* 3. 195–215. Available online at: <https://czasopisma.kul.pl/index.php/LingBaW/article/view/5659> (Accessed 2026-07-06).
- Viberg, Åke. 2019. Phenomenon-based perception verbs in Swedish from a typological and contrastive perspective. In Éric Gilbert (ed.), *Perception, perceptibilité et objet perçu. Approches inter-langues* [Syntaxe & Sémantique 20], 17–48. Caen: Presses Universitaires de Caen.
- Viberg, Åke. 2020. Contrasting semantic fields across languages. (Using the verbs of cutting and breaking as an example). In Renata Enghels, Bart Defrancq & Marlies Jansegers (eds.), *New approaches to contrastive linguistics. Empirical and methodological challenges*. [Trends in Linguistics 336], 265–312. Berlin: de Gruyter Mouton.
- Viberg, Åke. 2025. What kind of verb is TAKE? The English and Swedish verbs *take/ta* and their correspondences in other European languages. Unpublished manuscript.
- Viberg, Åke (to appear). Word meaning in comparative and contrastive linguistics. In Dirk Geeraerts & Dylan Glynn (eds.), *The Cambridge handbook of lexical semantics*. Cambridge: Cambridge University Press.
- Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.
- Wierzbicka, Anna. 1987. *English speech act verbs. A semantic dictionary*. New York: Academic Press.
- Wilkins, David P. & Deborah Hill. 1995. When “go” means “come”: Questioning the basicness of basic motion verbs. *Cognitive Linguistics* 6(2/3). 209–259.
- Windfuhr, Gernot & John R. Perry. 2010. Persian and Tajik. In Gernot Windfuhr (ed.), *The Iranian languages*, 416–544. Abingdon: Taylor & Francis Group.

Frequency dictionaries/lists

<i>Language</i>	<i>Size</i> <i>Million words</i>	<i>Reference</i>
Czech	20	Čermák, František & Michal Křen. 2011. <i>A frequency dictionary of Czech</i> . New York: Routledge.

English	100	Kilgariff, Adam. 1998. <i>BNC lists</i> . Available at: https://www.kilgariff.co.uk/bnc-readme.html (Accessed 2026-07-06).
Estonian	15	The Frequency List of Estonian Literary Language https://www.cl.ut.ee/ressursid/sagedused1/ (Accessed 2026-07-06).
Finnish	43	Wiktionary:Frequency lists/Finnish/Press data https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Finnish/Press_data (Accessed 2026-07-06).
French	23	Lonsdale, Deryle & Yvon Le Bras. 2009. <i>A frequency dictionary of French</i> . London: Routledge.
German	20	Jones, Randall L. & Erwin Tschirner. 2020, 2 ^{ed} . <i>A frequency dictionary of German</i> . London: Routledge.
Greek	47	Hellenic National Corpus: http://hnc.ilsp.gr/index.php?current_page=main&lang=en (Accessed 2026-07-06; data obtained from Maria Gavrilidou December 2009 from the version that existed at that time)
Hungarian	777	Oravecz, Csaba, Tamás Váradi & Bálint Sass. 2014. The Hungarian Gigaword Corpus. <i>Proceedings of LREC 2014</i> . http://hnc.nytud.hu (Accessed 2026-07-06).
Lithuanian	1	Andrius Utka, Sudarė. 2009. <i>Dažninis rašytinės lietuvių kalbos žodynas: 1 milijono žodžių morfologiškai anotuoto teksto pagrindu</i> . Kaunas: Vytauto Didžiojo universiteto leidykla. https://sitti.vdu.lt/wp-content/uploads/2019/06/Dazninis_zodynas.pdf (Accessed 2026-07-06).
Persian	10	Miller, Corey & Karineh Aghajanian-Stewart. 2017. <i>A Frequency Dictionary of Persian</i> . London: Routledge.
Russian	150	Sharoff, Serge, Elena Manskaya & James Wilson. 2013. <i>A frequency dictionary of Russian</i> . London: Routledge.
Spanish	20	Davies, Mark. 2006. <i>A frequency dictionary of Spanish</i> . London: Routledge.
Swedish	1	SUC 1997. <i>SUC 1.0. Stockholm Umeå Corpus</i> . Produced by Dept. of Linguistics, Umeå University and Dept. of Linguistics, Stockholm University. CD-Rom.

Turkish	51	Aksan, Yeşim, Mustafa Aksan, Ümit Mersinli & Umut Ufuk Demirhan. 2017. <i>A Frequency Dictionary of Turkish</i> . London: Routledge.
Welsh	11	Knight, Dawn, Steve Morris, Beth Tovey-Walsh, Tess Fitzpatrick, Laurence Anthony. 2020. <i>Yr Amliadur: Frequency Lists for Contemporary Welsh</i> (Version 1.0.0). Cardiff University. https://doi.org/10.17035/d2020.0120164107

Appendix

The identification of CAN and MUST

In several cases, it is difficult to decide which verbs correspond most closely across languages. To a certain extent there are detailed contrastive studies of the relevant verbs. When this is not the case, a resource that has been used is the Word Alignment Database (WAD) included as a tool in OPUS (Tiedeman 2012). When nothing else is said, data are taken from the Subtitles corpus. All 13 languages except Welsh are represented in OPUS.

This appendix will use CAN and MUST, which are described in Table 8 in Section 4, as an example how OPUS has been used. The modals are particularly suited as a test case since there is a wealth of detailed studies against which the result can be evaluated. The same type of analysis has been carried out on several of the nuclear verbs, when the correspondence has been unclear, but it is not possible to account for all such analyses in detail. I plan to make available appendices or articles about other semantic fields based on OPUS and the many other corpora that now are becoming freely available.

CAN. The Word Alignment Database takes a word form in one language as input and shows the word forms that appear in the translations according to a formula. Table App1:1 shows the output when the English word form *can* is used as input. There is a possibility to choose among source texts. The option “Subtitles 2018” was chosen. The result should be seen as tentative, since it should be compared to corpus-based studies carried out with manual coding like the one by Aijmer (2019) referred to in Section 4.

#	prob	cze	#	prob	ell	>>	#	prob	est	>>
84078	(0.14)	může  	195178	(0.21)	μπορώ  		88296	(0.27)	saa  	
			84376	(0.17)	μπορώ να  		30348	(0.12)	suuda  	
			141662	(0.16)	μπορείς  		19842	(0.08)	võid  	
			50503	(0.11)	μπορεί να  		20956	(0.08)	saan  	
			67496	(0.08)	μπορούμε  		20392	(0.07)	võib  	
#	prob	fin	>>	#	prob	fre	#	prob	ger	>>
338682	(0.44)	voi  		622830	(0.44)	peux  	378566	(0.47)	kann  	
25520	(0.06)	voin  		318466	(0.26)	peut  	161446	(0.24)	können  	
22518	(0.05)	voit  		108756	(0.10)	pouvez  	103404	(0.18)	kannst  	
20664	(0.05)	pysty  								
25684	(0.05)	saa  								
#	prob	hun	#	prob	lit	>>	#	prob	rus	>>
53508	(0.08)	lehet  	8036	(0.20)	gali  		274684	(0.38)	могу  	
40018	(0.07)	tud  	4292	(0.13)	galiu  		102750	(0.17)	можешь  	
40230	(0.07)	tudok  	1576	(0.06)	galime  		96580	(0.14)	может  	
45904	(0.06)	tudom  	1638	(0.05)	Gali  		80262	(0.14)	можем  	
			1584	(0.05)	Galiu  		46258	(0.08)	можете  	
#	prob	spa	>>	#	prob	swe	Translations of <i>could</i>			
529512	(0.30)	puedo  		555334	(0.63)	kan  	#	prob	est	
414236	(0.26)	puede  					18648	(0.15)	võib  	
258850	(0.17)	puedes  					9886	(0.11)	võiks  	
111688	(0.09)	pueden  					4788	(0.05)	saaks  	
85698	(0.07)	pueda  								

Table App1:1. The translations of English *can* in OPUS Word alignment (accessed 2026-07-06).

In Table App1:1, the languages are shown in alphabetical order, but the following comments follow the order in Table 8. (*ell* refers to *elleniká* Greek). German is represented by present tense forms of *können* inflected for person and number. In Swedish, the form *kan* dominates completely. This is the present tense of the Swedish cognate *kunna* ‘to be able’, which has the full set of endings of an ordinary verb. It should be noted that verbs are not inflected for person and number in Swedish. The French and Spanish translations contain forms inflected for person and number of the

verbs *pouvoir* and *poder*, which are ultimately derived from Classical Latin *posse* ‘be capable of’ (present *poteo* ‘I am capable’), see Cornillie et al. (2009), Kronning (2024). Greek is represented by the present tense of *boró* ‘I can’ (the citation form) in several persons (*boró* [μπορώ] 1SG, *borís* [μπορείς] 2SG, *borí* [μπορεί] 3SG, *borúme* [μπορούμε] 1PL). This verb is combined with the subjunctive particle *na* (να), which shows up in two cases in Table App1:1. When the verb has an epistemic meaning, only the 3SG form *borí* is used (Tsangalidis 2009). Present-day Slavic languages have a modal verb related to Old Church Slavonic *mošti* to express Possibility (Besters-Dilger et al. 2009: 178). In Czech, the most frequent translations are inflected forms of Czech *moci* (*můž*e PRS.3SG, *mohou* PRS.3PL) and in Russian of *moč’* (*mogu* [могу] PRS.1SG, *možet* [может] PRS.3SG, *možeš’* [можешь] PRS.2SG, *možem* [можем] PRS.1PL, *mogut* [могут] PRS.3PL). Lithuanian is represented by present tense forms of the verb *galėti* inflected for person (*gali* 2/3SG, *galiu* 1SG, *galime* 1PL). In Welsh, CAN is expressed with the verbs *gallu* and *medru*, which are said to be interchangeable, but the choice is subject to regional variation. The verbs can refer both to ability/physical possibility as well as epistemic possibility, but the construction tends to shift (King 2016: 249–255).

In the list of Estonian translations, three different verbs are represented *võima* ‘can’ (*võid* PRS.2SG, *võib* PRS.3SG); *saama* ‘get; can’ (*saan* PRS.1SG and the form used in the present negative *ei saa*); *suutma* ‘be able’ (*suud* used in present negation: *ei suud*). In the translations of *could*, which are shown only for Estonian, forms of *võima* dominate (*võib* PRS.3SG, *võiks* COND.PRS.3SG, NEG: *ei võiks*), but there is also one form of *saama* (the conditional present combined with negation: *ei saaks*). *Saama* and *võima* are combined with one of the two infinitives, the *da*-infinitive. In Finnish, various inflected forms of *voida* ‘can, to be able’ dominate, but two other verbs also are represented. The most frequent translation in Hungarian is the verb *tud*, whose basic meaning is ‘know’, as in (1). There is no information about frequency of the modal use.

(1) Hungarian (OPUS Subtitles)

- a. *nem tudom, ki maga, barátom, de remélem,*
 not know.PRS.DEF.1SG who you friend.POSS.1SG but hope.PRS.DEF.1SG
tud főzni
 know.2SG.FORMAL cook.INF
 ‘I don’t **know** who you are, friend, but I hope you **can** cook.’ (Internal)

- b. *most Paul nem tud játszani*
 now Paul NEG know.PRS.3SG play.INF
 ‘Paul **cannot** play now.’ (External)

Possibility can also be expressed with the suffix *-hAt* and with *lehet* (<‘be’ + *-hAt*) ‘be possible’. The latter is a frequent translation of *can* according to Table App1:1 but not as frequent as *tud*.

MUST. Even though *MUST* has been used to symbolize the major modal verb that expresses Necessity, *must* is rivalled by alternative expressions in English. As can be observed in Table 8 in Section 4 that shows the frequency in the frequency dictionaries, English *must* is less frequent than its correspondences in the other languages. This is obviously connected to the ongoing development of alternative expressions of obligation. For that reason, the searches with OPUS Word Alignment Database that are shown in Table App1:2 are based on Swedish *måste*. Like English *must*, *måste* has the same form in present and past tense.

#	prob	cze		#	prob	ell		#	prob	eng	> >
30638	(0.17)	musí	✓✗	62568	(0.27)	πρέπει	✓✗	86450	(0.28)	have to	✓✗
15686	(0.09)	muset	✓✗			να	✓✗	73366	(0.25)	must	✓✗
				11548	(0.09)	πρέπει	✓✗	42300	(0.16)	gotta	✓✗
						να	✓✗	42164	(0.15)	need to	✓✗
				5526	(0.06)	πρέπει	✓✗	19740	(0.08)	got to	✓✗
				5314	(0.06)	πρέπει	✓✗				
#	prob	est		#	prob	fin		#	prob	fre	> >
35156	(0.21)	pean	✓✗	74056	(0.30)	täytyy	✓✗	75400	(0.27)	dois	✓✗
28158	(0.18)	peame	✓✗	61076	(0.22)	pitää	✓✗	48782	(0.18)	doit	✓✗
28540	(0.18)	pead	✓✗					28224	(0.11)	faut	✓✗
18480	(0.12)	peab	✓✗					14944	(0.07)	devez	✓✗
								12510	(0.06)	devons	✓✗
#	prob	ger		#	prob	hun		#	prob	lit	
106348	(0.35)	muss	✓✗	123714	(0.30)	kell	✓✗	3628	(0.11)	reikia	✓✗
77824	(0.28)	müssen	✓✗					3226	(0.10)	turi	✓✗
33374	(0.14)	musst	✓✗					1740	(0.06)	turiu	✓✗
								1304	(0.05)	turime	✓✗

(continued)

#	prob	rus	>>	#	prob	spa		Europ	prob	eng	>>
32972	(0.16)	должен	✓✗	15792	(0.11)	debe	✓✗	69536	(0.42)	must	✓✗
33644	(0.15)	нужно	✓✗	8106	(0.06)	tiene que	✓✗	15266	(0.10)	need	✓✗
26338	(0.14)	должны	✓✗	8096	(0.06)	tengo que	✓✗	12630	(0.10)	must be	✓✗
8992	(0.09)	надо	✓✗					10082	(0.08)	have to	✓✗
7442	(0.08)	должна	✓✗					8378	(0.07)	need to	✓✗

Table App1:2. The translations of Swedish *måste* in OPUS Word alignment (accessed 2026-07-06).

Starting with English this time, we can see that the most frequent translation in Subtitles is *have to*. *Must* is the second most frequent alternative closely followed by *got to*, and *gotta* and *need to* are also relatively frequent. For comparison, the program was also run on texts from Europarl (bottom of right column in Table App1:2). In this text type, *must* was clearly dominant and more frequent than the combined frequencies of *need (to)* and *have to*. These observations are in accordance with the thorough study by Aijmer (2019), which shows that *must* is the main proponent for epistemic modality but is losing ground to *have to* and other semi-modals in the area of deontic modality and this on-going change is related to a cultural change towards the use of less authoritative and face-threatening language that appears to have parallels in some other European languages.

German is represented by inflected forms of *müssen*, which is inflected for both tense and person. French *devoir* and Spanish *deber* have developed from Latin *debere* ‘must’, which originally meant ‘I owe to / I am indebted to’ (Cornillie et al. 2009). Table App 1:2 shows inflected forms of these two verbs, but French also uses the impersonal *faut* and in Spanish, expressions based on *tener* ‘have’ are slightly more frequent than forms of *deber*. However, in Europarl only inflected forms of *deber* appear. Greek uses the impersonal verb *prepi* (πρέπει), which like *boró* ‘can’ is combined with the subjunctive particle *na* (να) and a main verb inflected for person and aspect. The Czech translations of *måste* represent inflected forms of the verb *muset*, which is etymologically related to ‘must’ in several other Slavic languages (Polish *musieć*, Belorussian *music’*) and to *may* and its cognates in Germanic languages. The major translation in Russian is an adjective *dolžen* that takes an infinitive as complement. The adjective is inflected for gender (*dolžen* [должен] m., *dolžna* [должна] f., *dolžno* [должно] n., *dolžny* [должны] pl.). All but the neuter form appear in App1:2. Tense is indicated with the copula (Zero in present tense). A frequent alternative is the impersonal predicate *nužno* (нужно) and the less frequent *nado*

(надо). In Welsh (not included in App1:2), MUST is expressed with a noun *rhaid* ‘need, necessity’, see (2). *Rhaid* can also be used with epistemic meaning but then appears in a different construction (King 2016: 265–267)

(2) Welsh (King 2016: 265)

(mae) *rhaid i ti fynd*
 is need for you go
 ‘You must go.’

Lithuanian uses inflected forms of the verb *turėti* ‘have’ as the major expression and this holds also for Europarl. There is also an impersonal expression *reikia*, which is the third person present form of *reikėti* ‘require’ (< *reika* ‘necessity’) and is combined with a dative subject. Estonian is represented by inflected forms of *pidama*. This verb means both ‘must’ and ‘hold’ and has a similar inflection except for indicative past (see Raag 2014: 317). The Finnish cognate *pitää* ‘hold’ can also express obligation (see below). The only translation of *måste* in Hungarian is *kell* which is counted as a verb but lacks agreement markers. The subject is optionally marked with a possessive suffix on the infinitive as in *Mennem kell* ‘must go’ [go.INF.POSS.1SG + *kell*]. It can also be represented by a noun or pronoun in dative form.

SUMMARY. CAN is realized as one of the 3-6 most frequent verbs in the majority of European languages and its frequency expressed as LN typically reaches above 8.0, which is superseded only by BE and HAVE. MUST is also expressed as a frequent verb and its LN usually reaches above 7.0, which is the cut-off point for the 20 most frequent verbs, but the picture is more complex than for CAN. One reason for that is that constructions with HAVE (or in Finnish BE) are frequent alternatives to a modal verb in several languages especially in certain text types.

Complementing the analysis with concordance data

The analysis can be taken one step further with another tool than WAD: OPUS – Corpus query (CWB), which allows searches in parallel corpora of a word in one language as input and provides concordances of the word as key word in context (KWIC) and translations into several other languages. Searches can be done i.a. in Subtitles and Europarl. Only a few short examples will be given here. This kind of analysis has not been carried out systematically due to the complexity of the semantics of the modals, but the tool has been used as a complement to the literature that has

been consulted. The tool is useful to show the characteristic constructions of the modals, especially when a language has several alternative correspondences, for example of Swedish *måste*. Example (3) shows the major correspondences in the languages which are displayed. All combine the modal verb with a main verb in the infinitive. In Finnish, the subject is in the genitive.

- (3) SWE Du *måste* förbli vaken, min son! (source: OPUS Subtitles)
 ENG You *must* stay awake, my son!
 FRE Tu *dois* rester éveillé.
 SPA *Debes* permanecer despierto, hijo.
 FIN Sinun *täytyy* pysyä hereillä!
 you.SG.GEN must.PRS.3SG stay.1INF awake

Example (4) shows the impersonal construction *il faut* in French and the use of *tener* ‘have’ to express MUST in Spanish, and the Finnish example in (4) shows the use of *pitää* ‘hold’ to express MUST in Finnish. Also in this case, the subject is marked with genitive.

- (4) SWE Vi *måste* förflytta oss snabbt. (source: OPUS Subtitles)
 we must move REFL.1PL fast
 ENG We’ve *gotta* move fast.
 FRE Il *faut* agir vite.
 it must.PRS.3SG act fast
 SPA *Tenemos* *que* ser rápidos.
 have.PRS.1PL that be fast.PL
 FIN Meidän *pitää* iskeä heti.
 we.GEN hold.PRS.3SG strike.1INF fast

In Finnish, there is also a special construction with *olla* ‘be’ combined with a passive present participle as in (5), which does not appear in Table App1:2.

- (5) Finnish (OPUS Subtitles)
 sinun on nai-ta-va Aella
 you.SG.GEN be.PRS.3SG marry-PASS-PRS.PTCP Aella
 ‘You must marry Aella.’

In Europarl, this alternative dominates completely, see (6) for a typical example.

(6) Finnish (Europarl)

taistelu on käy-tä-vä pääosin kansainvälise-llä taso-lla
 battle is employ-PASS-PRS.PTCP mainly international-AD level-AD
 ‘This battle must, in the main, be fought at international level.’

The corpus tool (CWB) can also be used to study the functions with which the modals are used and to test the requirement that modals should express no less than two types of modality. Example (3) above shows how the major modals in the languages that are included are used to express deontic modality (obligation). The following example show how they can also express epistemic modality. In (7), modality is expressed by an adverbial in Finnish, which is a rather frequent option also in the other languages. (Actually, it is difficult to find examples where a modal is found as a translation in all languages.)

(7) [There is a knock.] That **must** be Mulder. (source: OPUS Subtitles)

SWE Det **måste** vara Mulder.
 FRE Ce **doit** être Mulder.
 SPA **Debe** ser Mulder.
 FIN Tuo on **varmaan** Mulder. [That is surely Mulder]

Example (8) shows how the modal *täytyä* can express an epistemic meaning, whereas in this example an adverbial is used in French.

(8) They **must** have spotted you. (source: OPUS Subtitles)

SWE De **måste** ha sett er! [They must have seen you]
 FRE Ils vous ont **sûrement** repéré. [They have surely detected you]
 SPA **Deben** haberte seguido. [They must have followed you]
 FIN Heidän on **täytynyt** nähdä sinut.

Abbreviations

1 = 1 st person	GEN = genitive	PRS = present
1 _{INF} = 1 st infinitive	INF = infinitive	PTCP = participle
2 = 2 nd person	NEG = negation	REFL = reflexive
3 = 3 rd person	PASS = passive	SG = singular
AD = adessive	PL = plural	
DEF = definite conjugation	POSS = possessive	

References

- Aijmer, Karin. 2019. What is happening to *must* in Present-day English? A contrastive perspective on a declining modal auxiliary. In Harumi Sawada, Yoshio Nitta & Masa-aki Yamanashi (eds.), *Bamen to shutaisei/shukansei* [Situation and subjectivity], 387–402. Tokyo: Hituzi Syobo.
- Besters-Dilger, Juliane, Ana Drobnjaković & Björn Hansen. 2009. Modals in the Slavonic languages. In Björn Hansen & Ferdinand de Haan (eds.), *Modals in the languages of Europe*, 167–197. Berlin: Mouton de Gruyter.
- Cornillie, Bert, Walter De Mulder, Tine Van Hecke & Dieter Vermandere. 2009. In Björn Hansen & Ferdinand de Haan (eds.), *Modals in the languages of Europe*, 107–137. Berlin: Mouton de Gruyter.
- King, Gareth. 2016. *Modern Welsh. A comprehensive grammar*. London: Routledge.
- Kronning, Hans. 2024. Les verbes modaux *devoir* et *pouvoir*. In *Encyclopédie Grammaticale du Français, en ligne*. Available online at: <http://encyclogram.fr> (Accessed 2026-07-06).
- Raag, Raimo. 2014. *Estniska från grunden* [Finsk-ugriska läromedel FU 25]. Uppsala: Uppsala University.
- Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. Istanbul: European Language Resources Association (ELRA).
- Tsangalidis, Anastasios. 2009. Modals in Greek. In Björn Hansen & Ferdinand de Haan (eds.), *Modals in the languages of Europe*, 139–163. Berlin: Mouton de Gruyter.

Electronic sources

OPUS. <https://opus.nlpl.eu/> (Accessed 2026-07-06).

CONTACT

ake.viberg@lingfil.uu.se

Linguistic typology and the cognitive science of non-WEIRD societies: The role of memory types

ANTONIO BENÍTEZ-BURRACO¹, SIHAN CHEN² & DAVID GIL³

¹UNIVERSITY OF SEVILLE - ²MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE (USA) &

³INDEPENDENT SCHOLAR

Submitted: 04/10/2024

Revised version: 25/03/2025

Accepted: 27/11/2025

Published: 24/07/2026



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

Abstract

Linguistic typology has identified properties shared by the world's languages, as well as features with respect to which languages diverge, including infrequent or rare phenomena. Cognitive biases are one important source of language universals, through their indirect effect on language change via their impact on language learning and language use. At the same time, typologically rare phenomena can exist only if the human brain is able to accommodate them. Nonetheless, our knowledge of how we process language mostly derives from the analysis of a limited set of languages associated with WEIRD (Western Educated Industrialized Rich Democratic) societies. Including non-WEIRD languages in our psycholinguistic experiments stands to benefit the cognitive science of language and, in addition, linguistic typology, given that languages do best what speakers do most, which in turn depends on what our brains do most. In this paper, the benefits of a less biased approach to the cognitive science of language diversity are illustrated through the consideration of a research project aimed at determining the impact of the typological diversity of human languages on different memory types, specifically, procedural vs. declarative memory, with the latter encompassing two subtypes of relevance to linguistic structure, semantic and episodic. Our focus is on typological features that have been shown, in the past, to correlate with sociolinguistic factors. Overall, we aim at delving into an array of multi-directional feedback relationships among memory types, sociolinguistic properties, and typological features.

Keywords: cognitive biases; typological diversity; non-WEIRD languages; non-LOL languages; language complexity; memory types.

1. Introduction: giving a voice to non-WEIRD societies and their languages

Building on a long tradition of psychological and neurological studies of language, in recent years a cognitive science of language has emerged as a lively field of research, whose aim is to integrate diverse neuroscience approaches to language acquisition, processing, and even evolution. Not only is this approach expected to significantly improve our understanding of the nature of language, but in addition it stands to contribute significantly to our comprehension of human perception and cognition, and more generally, the human phenotype. For typological studies, the cognitive science of language will likely improve our understanding of the typological diversity of the world's languages, since cognitive biases contribute to determining what is frequent and what is rare among the world's languages. Nonetheless, research in the cognitive science of language still suffers from a dramatic sampling bias. Our manifesto article is intended to stress the benefits of including in our psycholinguistic and neurolinguistic research a larger number of languages, specifically those associated with non-WEIRD (i.e. Western, Educated, Industrialized, Rich, Democratic; Heinrich et al. 2010) societies, henceforth referred to simply as "non-WEIRD languages" – a kindred notion being that of "non-LOL languages" (where LOL stands for Literate, Official and with Lots of users; Dahl 2015). As an example of this less typologically-biased approach to language diversity, we present our ongoing research project, situated at the crossroads between typology, sociolinguistics, and psycholinguistics. The project is aimed at determining the potential impact on our cognition of aspects of the structural diversity of the world languages resulting in part from sociopolitical factors (although the opposite effect could be also expected, with changes in memory types also impacting on language structure). Specifically, we aim to test whether speakers of languages spoken in close-knit societies store linguistic information in the procedural and declarative memory systems in ways that differ from speakers of languages used by larger and more interconnected human groups.

The project thus aims to enrich extant hypotheses about the impact of language diversity on human cognition (via social environment-language-cognition co-evolution). As noted, for linguistic typology, knowing more about the cognitive differences shown by speakers of typologically diverse languages should improve our understanding of the nature of language universals and language *rara*.

Variation certainly pervades language. Language fulfils many functions, we speak thousands of languages that are structurally diverse, and different varieties of each

individual language can be found in different social groups, geographical areas, age groups, genders, contexts, and so on. Describing all this diversity has been a concern for many linguistic disciplines, ranging from dialectology through sociolinguistics to typology. While this is already a formidable task, it is even more challenging to provide an explanation for all this diversity. For a long time, typology has attempted to account for the diversity found among present-day languages in diachronic terms, pertaining to how language changed in the past. However, as noted by many other typologists, language change is also dependent on patterns on language acquisition and use, which are subject to diverse constraints many of which are of a cognitive nature. As noted by e.g. Culbertson (Culbertson 2012; Culbertson et al. 2013) these are typically biases favoring salience, harmony, or systematicity in language structure, which are frequent phenomena in the world's languages. Nonetheless, the world's languages also exhibit infrequent or even rare phenomena, which, as stressed by Cysouw and Wohlgemuth (2010), can tell us as much about the capacities and limits of human languages as do universals. For example, it might be suggested that typological *rara* are less learnable or more demanding in terms of cognitive resources. Knowing more about how our brain deals with the reduced cognitive demands of typological frequent phenomena or with the increased cognitive loads and demands of typological *rara* can help understand the typological diversity of human languages. That said, this cognitive approach to linguistic typology, even if sounded is not uncontroversial, and it certainly needs more evidence to be fully supported. Importantly too, non-cognitive factors are still much needed to explain typological patterns of interest. As noted by e.g. Harris (2010) regarding typological *rara*, they might exist despite their extra cognitive burden because of historical reasons, if they result from a large amount of otherwise common diachronic changes each demanding no extra cognitive effort.

For typologists, the structural diversity of human languages is certainly a gift. Diversity enables us to test assumptions about what is universal in languages, and what is specific to certain languages, language families, or linguistic areas. Nonetheless, as far as the cognitive underpinnings of all this diversity are concerned, we still suffer from a dramatic sampling bias, which resembles the sampling bias suffered by early linguistics. Specifically, psycholinguistic and neurolinguistic research still mostly focuses on a small set of languages, typically WEIRD languages, like English and other widespread Indo-European languages (see Blasi et al. 2022 and Benítez-Burraco 2025 for discussion), or what has been more recently called LOL

languages. Just as we already have detailed accounts of the typological diversity of human languages (like WALS, Dryer & Haspelmath 2013 or Grambank, Skirgård et al. 2023), so we also need richer accounts of how the human brain processes other types of languages. Fortunately, non-WEIRD/non-LOL languages are being increasingly examined by the neuroscience of language, as in the recent study by Malik-Moraleda and colleagues (2022) of syntax processing by speakers of 45 languages from 12 different language families. Ideophones are another notable example, as they are typically found in non-WEIRD languages. Research suggests that their cross-linguistic similarities might be due to their closer relationship (compared to non-ideophonic words) to motor and sensory components of our cognition, which are rather invariant among humans (De Varda & Strapparava 2022), with ideophones activating sensory areas in a greater degree and in a way that can be regarded as (quasi-)synesthetic (Osaka et al. 2004; Osaka 2009).

Both conceptually and methodologically, this research is expected to be challenging. Typological research of minority languages is difficult per se (see Gorter 1997; Nagy 2009; Marten et al. 2012, for interesting discussions). Conducting neurolinguistic and psycholinguistic research with speakers of these languages can be even more challenging. Technical facilities can be difficult to transport to and use in the testing places, tests need to be adapted, the necessary number of subjects for achieving statistically significant results might be not easily reached, and so on. In the next section, we provide a detailed description of our ongoing project examining some cognitive correlates of certain basic aspects of the typological diversity of languages. As part of our research, we are examining non-WEIRD languages, and we will also comment on potential ways of dealing with the methodological problems and limitations inherent to this type of research.

2. Cognitive biases, language diversity, and sociopolitical factors: a research project

2.1. Rationale and background

As discussed in the previous section, humans exhibit cognitive biases that favor specific ways of processing and storing linguistic information, and that are expected to contribute in some way to the existence of universal properties of languages, as found by typological research. If we further assume that human cognition (including

our faculty of language) is similar in all human beings (Moro 2008; Bolhius et al. 2014), one can certainly wonder why cross-linguistic differences do exist. One possible answer is, as noted above, historical change. Another possibility is that languages diversify due to their being host to different forces pulling in different directions, e.g. the need for expressive power as opposed to the drive for simplification. Finally, it has been suggested that cross-linguistic differences can also result from languages adapting to a changing environment. For instance, sound inventories might be, at least in part, a response to different physical environments, in view of the correlations found with vegetation coverage (Maddieson & Coupé 2015), humidity (Everett et al. 2015), or temperature (Wang et al. 2023). Likewise, the social environment might also exert some impact on language structure, this including the number of speakers, the degree of bilingualism, the tightness or looseness of the social networks, aspects of sociopolitical organization, or the number of adult learners of the language (Sinnemäki 2009; Lupyan & Dale 2010, 2016; Trudgill 2011; Nettle 2012; Atkinson et al. 2019; Gil 2021, among many others).

More specifically, it has been claimed that societies featuring large and complex social networks, involving greater rates of inter-group contacts and cultural exchanges (which are overwhelmingly WEIRD or LOL societies), might favor languages with larger lexicons, more complex phonologies, increased compositionality and enhanced semantic transparency, as well as more complex and more layered syntaxes, with more specialized and obligatory grammaticalized distinctions. Allegedly, these languages also tend to exhibit decreased complexity in some domains, particularly, their morphologies. Following an original proposal by Wray and Grace (2007), these languages have been called Type X languages, because of the eXoteric (i.e. outwards) nature of the societies speaking them (see Chen et al. 2024 for details). By contrast, the languages spoken by small and isolated human societies (which are overwhelmingly non-WEIRD/non-LOL societies), featuring tight networks and high proportions of native speakers tend to show larger sound inventories and more complex phonotactics, as well as opaquer morphologies, but reduced semantic transparency and compositionality (with an abundance of idioms and formulaic language), as well as simpler and less layered syntaxes. These languages have been called Type S languages, because of the eSoteric (i.e. inwards) nature of the societies speaking them (see Chen et al. 2024 for details).

The distinction between Type X and Type S languages is not uncontroversial. On the one hand, some authors have failed to find robust evidence of exoteric societies speaking languages that differ in some of the aspects mentioned above from the

languages spoken by esoteric societies (e.g. Shcherbakova et al. 2023 on morphology), hence more research is needed to properly support such a distinction. On the other hand, the distinction itself is quite crude. First, these two language types are merely ideal types (or poles) of a continuum of diversity, so that most if not all languages can be expected to show a complex admixture of type X and type S features (this is clearly observed in the statistical study by Chen et al. 2024). Second, the societal distinction supporting these two types of languages, namely, esoteric vs exoteric societal types, is itself a gross simplification. In actual fact, many or most small communities are not really "isolated", at the very minimum they may be multilingual, so even if outsiders do not typically learn their language, insiders do know the language(s) of other human groups. This is expected to impact on the features of the local (allegedly, Type S) language. Importantly, this circumstance may have a potentially significant impact on our project, which we consider in detail in section 2.3. below.

As discussed by several authors (e.g. Bolender 2007, Wray & Grace 2007, Trudgill 2011), the structural differences between Type X and Type S languages might result, to some extent, from their differential context-dependency. The reason is that speakers of the latter type constitute dense social networks and share substantial common ground, whereas users of Type X languages do not know each other to a similar extent and constitute looser social networks. Now, there is also increasing evidence that the languages we speak impact, even if subtly, on our cognition, this including our perceptive, processing, and storing abilities. For instance, speakers of languages with a S(ubject) V(erb) O(bject) basic order outscore speakers of languages with a SOV order at recalling final stimuli from a set or a list (and vice versa) (Amici et al. 2019). Likewise, it is well-known that languages differ in the degree to which their word-order patterns "harmonize". For example, while in Indonesian all noun modifiers follow the head noun, in English some noun modifiers, e.g. relative clauses, follow the head noun while others, e.g. adjectives, precede it. It has been observed that speakers of languages with less harmonic structures show a reduced tendency to regularity when dealing with structural rules, which is an important cognitive bias accounting for several typological findings (Culbertson 2012). Still, this does not mean that harmony or even regularity, as found in the world languages, are solely the outcome of our cognitive biases. Indeed, harmony might be favoured by some functional reasons. Nonetheless, research using computational simulations suggests that this is a more unlikely possibility than the cognitive reason (Chater et al. 2009).

It may be expected that given sufficient time, language-specific structural constraints of this type result in a differential involvement of specific cognitive abilities in language processing. In our project, we focus on two basic types of memories which play a central role in language processing: procedural and declarative. Procedural memory supports the acquisition and processing of compositional, rule-governed aspects of language; by contrast, declarative memory is typically involved in learning and storing words and irregular phenomena across domains, including suppletive forms, irregularities, and formulae (see Ullma 2004, 2015 for details). Typically, declarative memory encompasses two subtypes of memory: semantic and episodic (Tulving 1972). Whereas semantic memory is important for word components storage, episodic memory enables us to mentally move backwards and forwards in time, that is, remembering or foreseeing events. Intriguingly, episodic memory has been related to recursion, which is a typical feature of complex syntax (see Corballis 2019 for details). Because of the opposite features purportedly exhibited by Type X and Type S languages in the domain of grammar, we predict a greater involvement of procedural memory in the acquisition and processing of Type X languages, and, conversely, of declarative memory in the acquisition and processing of Type S languages. This is the main hypothesis we wish to test in the project (sketched in Chen et al. 2023), which is derived from the more general view that languages react adaptively to the sociopolitical environment in which they are spoken and that these changes are later cognitively accommodated by speakers via the potentiation of cognitive resources otherwise not specific to language that are needed for their acquisition and use, particularly, different types of memory. In truth, this suggested connection between memory types, specific domains of language, and different social contexts is not new. Among others, it has been proposed by Ullman himself, although in relation to second language learning (e.g. Ullman 2020). As noted by one anonymous reviewer, some experimental and typological work also builds on this idea (e.g., Clahsen et al. 2010; Sinnemäki & Di Garbo 2018).

In order to achieve our goals, we plan to test groups of 20-40 native speakers of strongly Type X and strongly Type S languages for procedural and declarative memory abilities. With regards to the latter, Type S languages, this means that we will be examining distinctively non-WEIRD/non-LOL languages. In this way, we expect to contribute to the improvement of a non-WEIRD cognitive science of language, as proposed in the first section of the paper.

That said, a non-trivial concern is the distinction between the two subcomponents of the declarative memory (episodic and semantic) which, as noted, can be expected to contribute differently to different aspects of language. One problem is that there is an ongoing, lively discussion about the ultimate nature of these two subtypes of memory, particularly, of semantic memory. Accordingly, for some researchers, semantic memory is the component of the declarative memory system (located in the hippocampal/medial temporal lobe) which facilitates the learning of “facts” about the world (in contrast to episodic memory, which supports the learning of the world’s “events”). Others, however, might use the term “semantic memory” to refer to the memory system that stores our knowledge about the world over the long term. Notice that this type of knowledge is sensitive to the properties of specific cultures. Because of this, most tests used to evaluate semantic memory tend to be culturally biased, which is an important concern for the type of cross-linguistic research we will be pursuing. These tests include the well-known Pyramids and Palm Trees test (PPTT) (Howard & Patterson 1992), but also the Camel and Cactus test (CCT) (Bozeat et al. 2000), which is a less categorically stringent version of the former. In the PPTT, participants are typically shown a target picture together with two other pictures: one semantically related to the target and another semantically unrelated. This procedure allows to evaluate semantic knowledge via non-verbal responses. The CCT is similarly designed and used, although participants are typically offered 4 possible images, with only one being semantically related to the target image. Nonetheless, even the modified version of the CCT test recently developed by Moore and colleagues (2022), which is claimed by its authors to be language-independent, exhibits some cultural biases.

A second important problem is that these two subcomponents of the declarative memory may exhibit some partial overlap and functional interconnection (see De Brigard et al. 2022 for a recent discussion). Learning and remembering new semantic knowledge is potentially inseparable from the specific event in which this knowledge was first presented, with events accordingly being dependent on episodic memory abilities. This seemingly explains why performance on the CCT test correlates well with brain regions associated with episodic memory (hippocampus) (Jana Reifegerste, personal communication).

The problems and limitations discussed above can be extended, in truth, to the two main memory types considered in our project, declarative and procedural. First, their neurobiological and cognitive status is still object of discussion (see e.g. Michaelian

2015 on declarative memory). More importantly, both memory systems show partly overlapping/redundant functions with regard to (specific components of) language.

Despite all these circumstances and potential limitations, we still believe that the distinctions above can be helpful in interpreting the typological results pertaining to language diversity, as well as the sociolinguistic findings about societal types (see Benítez-Burraco et al. 2022 for some reasons), among which are the purported trade-offs in complexity among different linguistic domains (of the sort discussed in e.g. the contributions in Sampson et al. 2009).

2.2. Methods

In order to remedy, or at least minimize all these caveats and limitations, we have decided to test our participants, specifically, for their ability to learn new items (as a proxy of their ability to learn words and the arbitrary, non-compositional aspects of language) or to learn new rules (as a proxy of their ability to learn the systematic aspects of language). In other words, we will be trying to distinguish, as much as possible, between their ability to learn new information and their ability to store such information at the long term, which is what most if not all available tests for semantic memory typically do. Of course, these two abilities can be expected to correlate: if you are good at learning new words, you will probably show a good long-term semantic memory; conversely, if you have managed to store a lot of semantic knowledge, this is probably because you first successfully learned it).

For testing procedural memory abilities, we will rely on an optimized version of the Serial Reaction Time (SRT) task (Nissen & Bullemer 1987), whereas for evaluating declarative memory abilities, we will use an optimized (and slightly adapted) version of the DecLearn task (Hedenius et al. 2013). Optimization here means, in essence, that we have developed improved versions of these tasks that can be taken online using a laptop or a mobile phone, but particularly, that we have carefully selected the stimuli to avoid any potentially confounding cultural effects, which, as noted, can be a limitation of most of the tests aimed to evaluate memory, particularly, semantic memory¹.

In the procedural memory task, a target stimulus (here: a green plus sign) appears in one of four horizontally aligned locations on the screen, which map onto four keys on the keyboard (S, F, J, L) (Figure 1). Participants are asked to press (using

¹ Development of these improved versions was conducted in collaboration with Jana Reifegerste.

their left and right middle and index fingers) the corresponding key as soon as the target stimulus appears on the screen. The task encompasses 6 blocks each of 60 trials each. In each trial, items remain on the screen until a correct response is registered, followed by a 100-ms inter-stimulus-interval, after which the next stimulus appears. Unbeknownst to the participant, the location of the target stimulus follows a 10-location sequence in some blocks (blocks 2-5), while it is random in other blocks (blocks 1 and 6). Previous research has shown that response times decrease during the sequence blocks, indicating that participants implicitly learn the sequence of locations. Participants will be encouraged to perform the task as quickly as possible. Responses by participants will be coded for both reaction time and accuracy. The dependent variable (participants' 'procedural memory score') is the procedural learning effect, which will be computed by first dividing the median reaction time of each block by the median reaction time of block 1, and then subtracting the resulting proportion in the last sequence block (i.e., block 5) from the proportion in the last random block (i.e., block 6) (see Juhasz et al. 2019 for details). Only correct responses will be considered.

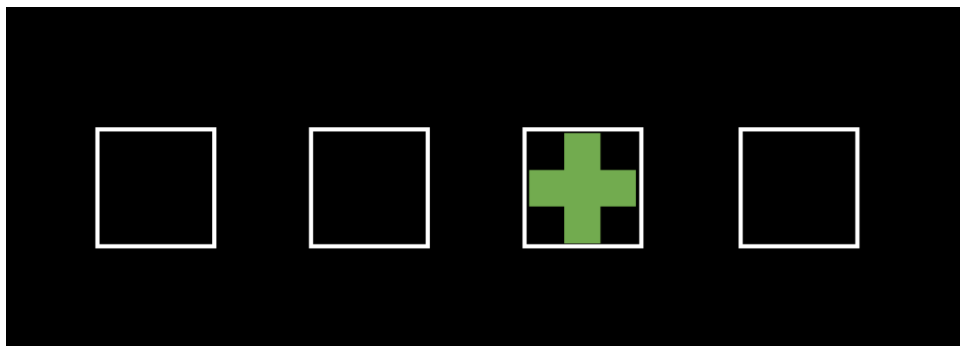


Figure 1. One of the stimuli used in the procedural memory task.

The declarative memory task encompasses two distinct phases, encoding and recognition. The first of these is a "dummy" phase, whose purpose is to expose participants to images that they will later be called upon to remember, but without clueing them in to the fact that their recollection of these images will be tested in the second, recognition phase. In order to avoid cross-cultural biases, the images are all abstract and non-representational, in order to ensure that participants have no prior familiarity with them. In the first, encoding phase, participants are presented with 60 triad tasks, each of which consists of a 'target' image in the top

centre of the screen and two horizontally arranged ‘foil’ images below. Participants are then asked to indicate (via key press on a computer keyboard; keys E and P) which of the two foil images more closely resembles the target. Each foil differs from the target in one out of 5 features: background, colour, number, orientation, or shape (e.g., target: three red circles inside a larger grey circle; foil A: three yellow circles inside a larger grey circle [colour condition]); foil B: one red circle inside a larger grey circle [number condition] (Figure 2A). Each trial consists of a fixation cross (1000 ms), item presentation (timeout: 8000 ms), and a blank screen (1000 ms). The five conditions are counterbalanced with regard to their location on the screen (left vs. right). The encoding phase comprises 50 experimental trials. As there is no objectively correct (vs. incorrect) response, an additional 10 “catch trials,” in which one of the two foils is very different from the target, serve to identify participants who do not follow the task instructions (e.g., by always pressing the same key). During the recognition phase, participants are presented with 120 images – half of which were presented as targets during the encoding phase, more specifically the larger images that appeared at the top of the screen (Figure 2B), while the other half were not presented (Figure 2C) –, and are asked to indicate via key press on a keyboard which images they recognize as having seen during the encoding phase. The dependent variable (participants’ ‘declarative memory score’) is their d-prime (d') score in the recognition phase, which will be computed as the z-score of the hit rate (previously presented items correctly identified as having-been-seen) minus the z-score of the false alarm rate (items that were not presented incorrectly judged as having-been-seen).

Participants will be never informed that they are being tested on their memory for triads of objects, with the aim of minimizing the involvement of working memory (see Logan et al. 2002; Takashima et al. 2006, or Craik & Rose 2012 for discussion).

Notice that our declarative memory test is also a nonverbal task. This has some advantages on its own. First, it facilitates more direct and reliable comparisons across speakers of different languages, since words are more culturally dependent, as noted. Second, the control for psycholinguistic properties of the stimuli becomes simplified, since one does not need to consider aspects like frequency or length, as in the case of using words. At the same time, the non-verbal nature of the DecLearn task also entails that what it truly evaluates is episodic memory abilities, since it assesses whether participants can remember if they have encountered a given image during a previous

phase of the task. However, for the reasons discussed above, we expect the results in this test to be a reliable proxy for their semantic memory abilities too, even if this ability is understood as the ability to store words in the long term. In fact, there is evidence that the results in the DecLearn task significantly correlate with the results obtained in the CCT test (Jana Reifegerste, personal communication).

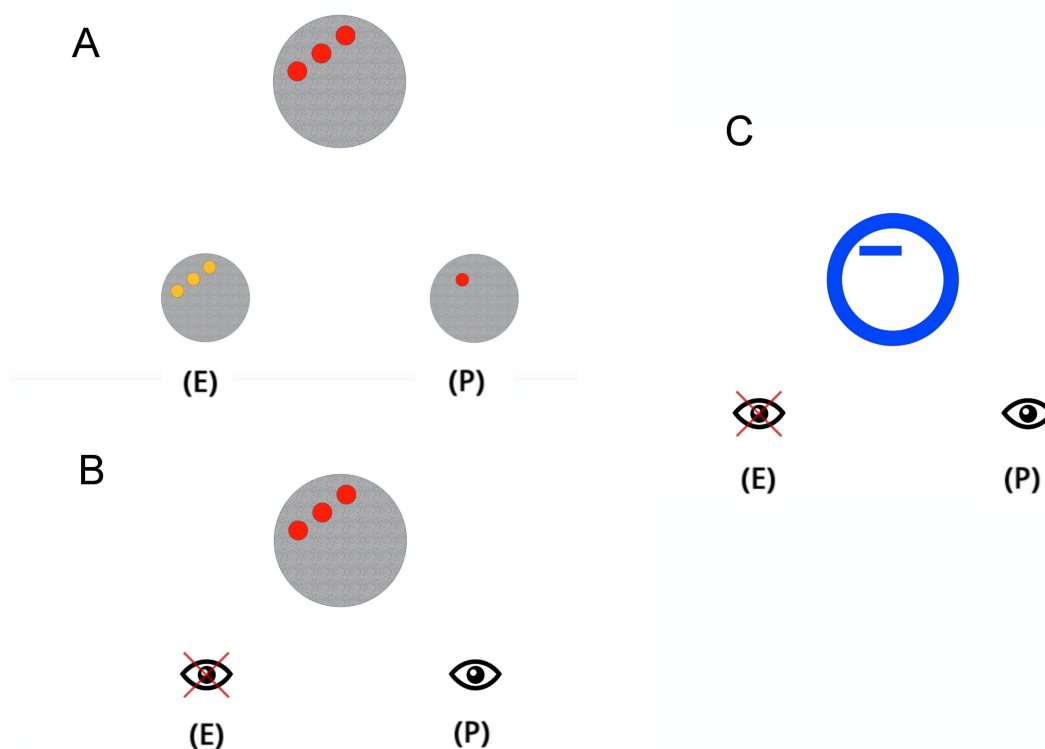


Figure 2. Examples of the stimuli used in the declarative memory task. A) One of the triads encompassing the encoding phase of the task. B) One of the triads in the recognition phase which include items that were seen during the encoding phase. C) One of the triads in the recognition phase which do not include items previously seen during the encoding phase. In all cases, (E) and (P) stand for the keys E and P in the keyboard of the device used for passing the task.

As noted, the distinction between Type X languages and Type S languages, although gross and pending of further confirmation, is expected to have some sort of sociopolitical motivation; moreover, it involves diverse language domains (phonology, morphology, syntax, and the lexicon). In our project we will be focusing on morphology and syntax only. That said, linking Type S languages (and particularly, a rich morphology) to an enhanced declarative memory (or more specifically, to an enhanced semantic memory), and Type X to an enhanced procedural memory (and

perhaps to an enhanced episodic memory) might also be a crude simplification. One basic reason is that all these types of memories are needed for processing any aspect of language. For instance, morphology is also rule-dependent, which means that procedural memory is involved in word formation too. Moreover, morphology can also, albeit in rare cases, exhibit recursion, which means that episodic memory is necessary for morphology too. Likewise, recursion involves rules, and rules are stored in the procedural memory. For these reasons, we have also planned to perform a second round of psycholinguistic analyses, aimed to test these memory abilities on groups of 20-40 native speakers of languages with complex morphology + complex syntax vs languages with simple morphology + simple syntax.

Additionally, it is important to bear in mind that these types of memory are “domain-general” in the sense that they underlie multiple cognitive domains. For instance, working on an assembly line typically entails mastering procedural abilities that are expected to be supported by a potentiated procedural memory. Still, this circumstance does not preclude sub-specialization of these memory types for language, as established with evidence from other domains. It is just that other factors external to language need to be properly controlled for matching subject groups in order to obtain conclusions that are informative about language specifically. Incidentally, one could hypothesize that changes in these types of memories in response to factors external to language might favor (or disfavor) Type X or Type S features in the languages spoken by various human groups. Accordingly, if people routinely engage in activities potentiating procedural memory abilities, they could find Type X languages more learnable, thus favoring their distinctive features over Type S features. And vice versa. Finally, we have planned to test speakers of the lowest socioeconomic status, in order to minimize the impact of literacy (and accordingly, the effect of the more exoteric varieties of their native languages), as well as the exposure to widespread Type X languages, like English.

2.3. Contingencies

The potential difficulties and limitations of our planned research do not differ substantially from those experienced by most field studies of minority languages and more specifically, by psycholinguistic research on such languages (see Adamou 2023 for a recent review). One critical point will be testing a suitable number of native speakers of Type S languages, which are spoken in remote places that are often difficult to access.

Also, the necessary facilities for testing speakers might be absent or difficult to provide (laptops, power supply, internet connection, and the like). Additionally, speakers are likely not to be particularly skilled in using most of these facilities, particularly, laptops and online resources, often requiring someone with the needed expertise to supervise the procedure on site. For all these reasons, the experiments will be always run in collaboration with field researchers who are familiar with the local languages. In addition, however, we will also develop online versions of the memory tests to be used for evaluating speakers via massive online calls or through platforms like Amazon Mechanical Turk (MTurk) or Prolific. Even if online tests may *prima facie* be regarded as more appropriate for speakers of WEIRD languages, currently, speakers of non-WEIRD languages can be found in many big cities in the countries where the languages are spoken. For instance, many speakers of minority Siberian languages (one of the groups we aim to test) live in Novosibirsk. Many of them still retain a good command of their mother tongues and have a reduced knowledge of Russian. Because of sociological and historical reasons, many of them can be (semi)illiterate. With the aid of speakers of their languages who are familiar with modern technologies, like smartphones or laptops (typically, younger people), such speakers can be properly tested using online tasks. Admittedly, this option is worse than testing speakers on site, but it can be good enough in cases where the access to local communities is particularly difficult.

A second critical point concerns the design of the tests. Most psycholinguistic tests are designed for being used with representatives of WEIRD societies, but the involved stimuli and procedures can be strange (or plainly incomprehensible) for speakers of minority, non-WEIRD languages. Cultural effects can be very subtle. The original version of the DecLearn and SRT memory tasks we will be employing made use of images depicting real and imaginary objects that different cultures might be familiar with to different degrees, thereby rendering the results incommensurable. In order to address this particular problem, we will adapt, as noted, both tests by replacing the original images with ones of an abstract, non-representational nature. At first, we decided to adapt the stimuli to the cultural context, but we experienced many problems for properly achieving this (mostly, because any stimuli that came to our minds could be argued to suffer from some kind of cultural bias); since we plan to use our tests with many different cultural groups, we finally decided to keep the stimuli as abstract as possible.

A third, major challenge faced by our research program stems from the fact that a large majority of the world's population today is multilingual and/or multiglossic. Thus, the results that we obtain from the experiment concerning the procedural and declarative

memory capabilities of particular participants cannot be linked in straightforward one-to-one fashion to particular languages and their associated grammatical characteristics; instead, it is necessary to take into account the entire linguistic ecology associated with individual participants. This problem is particularly acute in the case of the non-WEIRD Type X languages that are of greatest interest to this project. To cite just one example, hunter-gatherers from the island of Mentawai, off the west coast of Sumatra, in Indonesia, will typically enjoy varying degrees of mastery of — in order of increasing exotericity — the indigenous Mentawai language, the regional lingua franca Minangkabau, the local colloquial koiné variety of Indonesian, and the national language Standard Indonesian. Thus, in order to make sense of our results, we intend to develop a method that will take such linguistic ecologies into account, by assigning appropriate weightings to each constituent language, in order to assign individual participant an overall grammatical and social profile.

Finally, an important limitation of our project is that it largely depends on distinctions (Type X language vs Type S language, exoteric societies vs esoteric societies) that, as discussed in detail in sections 2.1 and 2.2 above, are in need of further confirmation and refinement, as well as on a neurobiological/cognitive hypotheses about language processing that also needs to be fully established. This is why we consider our research as a pilot project, a preliminary exploration of the complex interfaces of linguistic typology, cognitive science, and sociolinguistics.

3. Conclusions

Overall, our project aims to shed light on the typological features, cognitive underpinnings, and sociocultural environments of non-WEIRD languages. More generally, we expect to contribute to current typological studies at the crossroads of classical typology, psycholinguistics, and sociolinguistics. In particular, we expect the project to contribute to ongoing discussions of the effects of the socio-cultural environment on language diversity, the nature of human communicative interactions (with a focus on the impact of the context on verbal structure), the role of different memory types (procedural and declarative) in language processing, and ultimately, the foundations of the typological differences between WEIRD and non-WEIRD languages.

Acknowledgements

This research was supported by grant PID2023-147095NB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

References

- Adamou, Evangelia. 2023. Experimental methods to study cultural differences in linguistics. In Sandrine Zufferey & Pascal Gygax (eds.), *The Routledge Handbook of Experimental Linguistics*, 458-472. London: Routledge.
- Amici, Federica, Álex Sánchez-Amaro, Carla Sebastián-Enesco, Trix Cacchione, Matthias Allritz, Juan Salazar-Bonet & Federico Rossano. 2019. The word order of languages predicts native speakers' working memory. *Scientific Reports* 9(1). 1124.
- Atkinson, Mark, Gregory J. Mills & Kenny Smith. 2019. Social group effects on the emergence of communicative conventions and language complexity. *Journal of Language Evolution* 4(1). 1-18.
- Benítez-Burraco, Antonio. 2025. The cognitive science of language diversity: achievements and challenges. *Cognitive processing* 26(3). 719–736.
- Benítez-Burraco, Antonio, Candy Cahuana, Sihan Chen, David Gil, Ljiljana Progovac, Jana Reifegerste & Tatiana Tatarinova. 2022. Cognitive and genetic correlates of a single macro-parameter of crosslinguistic variation. *Proceedings of JCoLE (Joint Conference on Language Evolution)*, 78-80. Kanazawa, Japan.
- Blasi, Damián E., Joseph Henrich, Evangelia Adamou, David Kemmerer & Asifa Majid. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences* 26(12). 1153–1170.
- Bolender, John. 2007. Prehistoric cognition by description: a Russellian approach to the upper paleolithic. *Biology & Philosophy* 22. 383-399.
- Bolhuis, Johan J., Ian Tattersall, Noam Chomsky & Robert C. Berwick. 2014. How could language have evolved? *PLoS Biology* 12(8). e1001934.
- Bozeat, Sasha, Matthew A. Lambon Ralph, Karalyn Patterson, Peter Garrard & John R. Hodges. 2000. Non-verbal semantic impairment in semantic dementia. *Neuropsychologia* 38(9). 1207–1215.
- Chater, Nick, Florencia Reali & Morten H. Christiansen. 2009. Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences of the United States of America* 106(4). 1015–1020.

- Chen, Sihan, David Gil, Sergey Gaponov, Jana Reifegerste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac & Antonio Benítez-Burraco. 2023. Linguistic and memory correlates of societal variation: A quantitative analysis. *PsyArXiv*. <https://doi.org/10.31234/osf.io/bnz2s>.
- Chen, Sihan, David Gil, Sergey Gaponov, Jana Reifegerste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac & Antonio Benítez-Burraco. 2024. Linguistic correlates of societal variation: A quantitative analysis. *PloS One* 19(4). e0300838.
- Clahsen, Harald, Claudia Felser, Kathleen Neubauer, Mikako Sato, Renita Silva. 2010. Morphological structure in native and non-native language processing. *Language Learning* 60(1). 21–43.
- Corballis, Michael C. 2019 Language, memory, and mental time travel: an evolutionary perspective. *Frontiers in Human Neuroscience* 13. 217.
- Craik, Fergus I. M. & Nathan S. Rose. 2012. Memory encoding and aging: A neurocognitive perspective. *Neuroscience and Biobehavioral Reviews* 36. 1729–1739.
- Culbertson, Jennifer. 2012. Typological universals as reflections of biased learning: evidence from Artificial Language Learning. *Language and Linguistics Compass* 6(5). 310–329.
- Culbertson, Jennifer, Paul Smolensky & Colin Wilson. 2013. Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science* 5(3). 392–424.
- Cysouw, Michael & Jan Wohlgemuth. 2010. The other end of universals: theory and typology of rara. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking Universals: How Rarities Affect Linguistic Theory*, 1-10. Berlin / New York: Walter de Gruyter.
- Dahl, Östen. 2015. How WEIRD are WALS languages? [Conference presentation]. The Diversity Linguistics: Retrospect and Prospect conference. May 1-3. Leipzig, Germany.
- De Brigard, Felipe, Sharda Umanath & Muireann Irish. 2022. Rethinking the distinction between episodic and semantic memory: Insights from the past, present, and future. *Memory & cognition* 50(3). 459–463.
- De Varda, Andrea Gregor & Carlo Strapparava. 2022. A cross-modal and cross-lingual study of iconicity in language: insights from deep learning. *Cognitive Science* 46(6). e13147.
- Everett, Caleb, Damian E. Blasi & Seán G. Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences USA* 112(5). 1322-1327.

- Gil, David. 2021. Tense-aspect-mood marking, language-family size and the evolution of predication. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 376(1824). 20200194.
- Gorter, Durk. 1997. Social surveys of minority language communities. In Brunon Synak & Tomasz Wicherkievicz (eds.), *Language Minorities and Minority Languages in the Changing Europe*, 59-76. Gdansk: Wydawnictwo Uniwersytetu Gdanskiego.
- Harris, Alice 2010. Explaining typologically unusual structures: the role of probability. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking Universals: How Rarities affect Linguistic Theory*, 91-103. Berlin: De Gruyter Mouton.
- Hedenius, Martina, Michael T. Ullman, Margareta Jennische Per Alm & Jonas Persson. 2013. Enhanced recognition memory after incidental encoding in children with developmental dyslexia. *PloS One* 8(5). e63998
- Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. The weirdest people in the world?. *The Behavioral and brain sciences* 33(2-3). 61–135.
- Howard, David & Karalyn Patterson. 1992. *Pyramids and Palm Trees: A test of semantic access from pictures and words*. Bury St. Edmunds, UK: Thames Valley Test Company.
- Juhász, Dora, Dezsó Nemeth & Karolina Janacsek. 2019. Is there more room to improve? The lifespan trajectory of procedural learning and its relationship to the between- and within-group differences in average response times. *PloS One* 14(7). e0215116.
- Logan, Jessica M., Amy L. Sanders, Abraham Z. Snyder, John C. Morris & Randy L. Buckner. 2002. Under-recruitment and nonselective recruitment: Dissociable neural mechanisms associated with aging. *Neuron* 33. 827–840.
- Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS One* 5(1). e8559.
- Lupyan, Gary & Rick Dale. 2016. Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences* 20. 649-660.
- Maddieson, Ian & Christophe Coupé. 2015. Human spoken language diversity and the acoustic adaptation hypothesis. *The Journal of the Acoustical Society of America* 138. 1838-1838.
- Malik-Moraleda, Saima, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev & Evelina Fedorenko. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience* 25. 1014-1019.

- Marten, Heiko F., Luk Van Mensel & Durk Gorter. 2012. Studying minority languages in the linguistic landscape. In Durk Gorter, Heiko F. Marten & Luk Mensel (eds.), *Minority languages in the linguistic landscape*, 1-15. London: Palgrave Macmillan UK.
- Michaelian, Kourken. 2015. Opening the doors of memory: is declarative memory a natural kind?. *Wiley interdisciplinary reviews. Cognitive science* 6(6). 475–482.
- Moore, Katrina, Rhian Convery, Martina Bocchetta, Mollie Neason, David M. Cash, Caroline Greaves, Lucy L. Russell, Mica T. M. Clarke, Georgia Peakman, John van Swieten, Lize Jiskoot, Fermin Moreno, Myriam Barandiaran, Raquel Sanchez-Valle, Barbara Borroni, Robert Laforce Jr, Marie-Claire Doré, Mario Masellis, Maria Carmela Tartaglia, Caroline Graff, Daniela Galimberti, James B. Rowe, Elizabeth Finger, Matthis Synofzik, Hans-Otto Karnath, Rik Vandenberghe, Alexandre de Mendonça, Carolina Maruta, Fabrizio Tagliavini, Isabel Santana, Simon Ducharme, Chris Butler, Alex Gerhard, Johannes Levin, Adrian Danek, Markus Otto, Jason D. Warren, Jonathan D. Rohrer & Genetic FTD Initiative, GENFI. 2022. A modified Camel and Cactus Test detects presymptomatic semantic impairment in genetic frontotemporal dementia within the GENFI cohort. *Applied neuropsychology. Adult* 29(1). 112–119.
- Moro, Andrea. 2008. *Impossible Languages*. Cambridge: MIT Press.
- Nagy, Naomi. 2009. The challenges of less commonly studied languages. In James N. Stanford & Dennis R. Preston (eds.), *Variation in indigenous minority languages*, 397-417. Amsterdam: John Benjamins.
- Nettle, Daniel. 2012. Social scale and structural complexity in human languages. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367. 1829–1836.
- Nissen, Mary Jo & Peter Bullemer. 1987. Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology* 19. 1–32
- Osaka, Naoyuki. 2009. Walk-related mimic word activates the extrastriate visual cortex in the human brain: an fMRI study. *Behavioural brain research* 198(1). 186-189.
- Osaka, Naoyuki, Mariko Osaka, Masanao Morishita, Hirohito Kondo & Hidenao Fukuyama. 2004. A word expressing affective pain activates the anterior cingulate cortex in the human brain: an fMRI study. *Behavioural Brain Research* 153(1). 123-127.
- Sampson, Geoffrey, David Gil & Peter Trudgill. 2009. *Language complexity as an evolving variable* (Vol. 13). Oxford: Oxford University Press.
- Shcherbakova, Olena, Susanne Maria Michaelis, Hannah J. Haynie, Sam Passmore, Volker Gast, Russell D. Gray, Simon Greenhill, Damián E. Blasi, & Hedvig Skirgård.

2023. Societies of strangers do not speak less complex languages. *Science advances* 9(33). eadf7704.
- Sinnemäki, Kaius. 2009. Complexity in core argument marking and population size. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 126-140. Oxford: Oxford University Press.
- Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: a typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. 1141.
- Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latache, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bower, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye 葉靖婷, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). eadg6175.
- Takashima, Atsuko, Ole Jensen, Robert Oostenveld, Eric Maris, Mara van de Coevering & Guillén Fernández. 2006. Successful declarative memory formation is associated

- with ongoing activity during encoding in a distributed neocortical network related to working memory: A magnetoencephalography study. *Neuroscience* 139. 291–297.
- Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.
- Tulving, Endel. 1972. Episodic and semantic memory. In Endel Tulving & Wayne Donaldson (eds.), *Organization of memory*, 381–403. Cambridge: Cambridge University Press.
- Ullman, Michael T. 2020. The declarative/procedural model: A neurobiologically motivated theory of first and second language 1. In Bill VanPatten & Jessica Williams (eds.), *Theories in Second Language Acquisition: An Introduction* (3rd ed.), 128–161. London: Routledge.
- Ullman, Michael T. 2015. The declarative/procedural model: a neurobiologically motivated theory of first and second language. In Bill VanPatten & Jessica Williams (eds.), *Theories in Second Language Acquisition: An Introduction* (2nd ed.), 135–158. London: Routledge.
- Ullman, Michael T. 2004. Contributions of memory circuits to language: The declarative/procedural model. *Cognition* 92(1–2). 231–270.
- Dryer, Matthew S. & Martin Haspelmath 2013. WALS Online. <https://doi.org/10.5281/zenodo.13950591>.
- Wang, Tianheng, Søren Wichmann, Quansheng Xia & Qibin Ran. 2023. Temperature shapes language sonority: Revalidation from a large dataset. *PNAS Nexus* 2(12). 384. <https://doi.org/10.1093/pnasnexus/pgad384>
- Wray, Allison & George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117. 543–578.

CONTACT

abenitez8@us.es

sihanc@mit.edu

dapiiiit@gmail.com

Corrigendum. Is complementation a universal strategy? A cross-linguistic corpus study

NICHOLAS EVANS^{1,2}, DANIELLE BARTH^{1,2}, WAYAN ARKA^{1,2,3}, HENRIK BERGQVIST⁴,
CHRISTIAN DÖHLER⁵, SONJA GIPPER⁶, YUKINORI KIMOTO⁷, DOMINIQUE KNUCHEL⁸,
DANIEL MAJCHRZAK⁹, HITOMI ŌNO¹⁰, EKA PRATIWI¹¹, SASKIA VAN PUTTEN¹²,
ANDREA C. SCHALLEY¹³, ASAKO SHIOHARA¹⁴, STEFAN SCHNELL¹⁵, YANTI¹⁶

¹SCHOOL OF CULTURE, HISTORY & LANGUAGE, AUSTRALIAN NATIONAL UNIVERSITY; ²CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE, AUSTRALIAN NATIONAL UNIVERSITY; ³UNIVERSITAS UDAYANA; ⁴DEPARTMENT OF PHILOSOPHY, LINGUISTICS, AND THEORY OF SCIENCE, UNIVERSITY OF GOTHENBURG; ⁵BERLIN-BRANDENBURG ACADEMY OF SCIENCES AND HUMANITIES; ⁶DEPARTMENT OF LINGUISTICS, UNIVERSITY OF COLOGNE; ⁷GRADUATE SCHOOL OF HUMANITIES, OSAKA UNIVERSITY; ⁸DEPARTMENT OF LINGUISTICS, UNIVERSITY OF BERN; ⁹SCHOOL OF LITERATURE, LANGUAGES AND LINGUISTICS, AUSTRALIAN NATIONAL UNIVERSITY; ¹⁰REITAKU UNIVERSITY; ¹¹FACULTY OF FOREIGN LANGUAGES, UNIVERSITAS MAHASARASWATI DENPASAR; ¹²CENTRE FOR LANGUAGE STUDIES, RADBOUD UNIVERSITY NIJMEGEN; ¹³DEPARTMENT OF LANGUAGE, LITERATURE AND INTERCULTURAL STUDIES, KARLSTADT UNIVERSITY; ¹⁴RESEARCH INSTITUTE FOR LANGUAGES AND CULTURES OF ASIA AND AFRICA, TOKYO UNIVERSITY OF FOREIGN STUDIES; ¹⁵INSTITUTE FOR THE INTERDISCIPLINARY STUDY OF LANGUAGE EVOLUTION, UNIVERSITY OF ZURICH; ¹⁶ATMA JAYA CATHOLIC UNIVERSITY OF INDONESIA



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

In the article:

Evans, Nicholas & Danielle Barth, Wayan Arka, Henrik Bergqvist, Christian Döhler, Sonja Gipper, Yukinori Kimoto, Dominique Knuchel, Daniel Majchrzak, Hitomi Ono, Eka Pratiwi, Saskia van Putten, Andrea C. Schalley, Asako Shiohara, Stefan Schnell, Yanti. 2026. Is complementation a universal strategy? A cross-linguistic corpus study. *Linguistic Typology at the Crossroads*, Vol. 5 No. 2 (2025), pp. 66–104.

<https://doi.org/10.60923/issn.2785-0943/21288>

one of the authors, Stefan Schnell, was inadvertently omitted from the list of authors in the published version of the article. This omission originated in the non-anonymous

version of the manuscript submitted after peer review and was not identified during copyediting and proof correction.

The error was amended on **3 March 2026**, and the PDF version available on the journal website has been replaced with a corrected version that includes Stefan Schnell in the list of authors. The scientific content of the article remains unchanged.

The authors regret this omission and apologize for any inconvenience caused.

References

Evans, Nicholas & Danielle Barth, Wayan Arka, Henrik Bergqvist, Christian Döhler, Sonja Gipper, Yukinori Kimoto, Dominique Knuchel, Daniel Majchrzak, Hitomi Ono, Eka Pratiwi, Saskia van Putten, Andrea C. Schalley, Asako Shiohara, Stefan Schnell, Yanti. 2026. Is complementation a universal strategy? A cross-linguistic corpus study. *Linguistic Typology at the Crossroads*, Vol. 5 No. 2 (2025), pp. 66–104. <https://doi.org/10.60923/issn.2785-0943/21288>

CONTACT

nicholas.evans@anu.edu.au