Linguistic Typology at the Crossroads

ISSN 2785-0943
Volume 4 – issue 2 – 2024
Issue DOI: https://doi.org/10.6092/issn.2785-0943/v4-n2-2024

Publisher
Department of Classical Philology and Italian Studies (University of Bologna)
Department of Modern Languages, Literatures and Cultures (University of Bologna)
The journal is hosted and maintained by **AlmaDL**

Linguistic Typology at the Crossroads

Responsible Editor

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

Linguistic Typology at the Crossroads

CONTENTS

Linguistic Typology at the Crossroads 4-2 (2024): 1-71 ISSN 2785-0943

A law of meaning

BERNHARD WÄLCHLI & ANNA SJÖBERG
STOCKHOLM UNIVERSITY, DEPT. OF LINGUISTICS

Submitted: 13/01/2024 Revised version: 28/10/2024
Accepted: 28/10/2024 Published: 23/01/2025

Abstract

This article rejects the canonical ideal of a one-to-one correspondence between meaning and marker and proposes a set-theoretical and optimality-based law for the relationship between meaning and its markers which allows for distinguishing true markers (such as *not*, *no*, *never* for negation) from otherwise associated items (such as negative polarity items as *but*, *any*): *a meaning is expressed by the set of non-randomly recurrent markers that together are the best collocation of that meaning*.

We implement the law in an algorithm using Dunning's log-likelihood and illustrate it by extracting markers for 'know', negation, first person subject and complementizers from translations of the New Testament in a variety sample of 83 languages with manual evaluation of all extracted markers. Markers are extracted from unannotated texts (with lexemes being just a special case of marker-set coalition phenomena) considering just one meaning at a time (without any need for accounting for specific coexpression types).

Keywords: semantics; parallel texts; collocation; optimality; negation; knowledge predication; personal pronouns and indexes; complementizers

1. Introduction

This paper proposes a general and explicit solution to the question of how to determine, in the absence of expert intuition and using a distributional approach, *by*

which markers a specific meaning (lexical or grammatical) is expressed in any language.[1] The importance of this question is downplayed in many approaches to linguistics, which – explicitly or implicitly – assume a canonical ideal of an exact one-to-one correspondence between meaning and marker, an assumption which we entirely reject. There are usually some contexts where a meaning is not expressed at all by a marker, markers usually express more than one meaning and a meaning is often expressed by more than one marker in a language; put differently, the correspondence is hardly ever one-to-one and there is hardly ever any exact congruence between one meaning and one marker.

We will demonstrate our distributional approach by using translations of the New Testament, a massively parallel text (Mayer & Cysouw 2014), in a genealogically and areally stratified sample of languages, considering the meanings 'know', negation, first person singular subject and propositional complementizers. Illustrating the general task with 'know' and French (fra; Indo-European, Romance), the task is to identify forms such as the ones in boldface in (1), *without* any expert intuition and thus without knowing anything about how French word-forms group to lexemes or that French has two verbs *savoir* 'know (fact)' and *connaître* 'know (person)'. All we have are parallel texts aligned roughly at sentence level; put differently, we have for each sentence translations in other languages.

(1) French; fra-x-bible-darby (40024033, 01044015, 41014071, 41012012)
   a. ***Sachez*** *que cela est proche, à la porte.*
   b. *Ne* ***savez-vous*** *pas qu'un homme tel que moi* ***sait*** *deviner?*
   c. *Je ne* ***connais*** *pas cet homme dont vous parlez.*
   d. *Ils* ***connurent*** *qu'il avait dit cette parabole contre eux*.

We approach the task by applying the notion of collocation, which can be broadly defined as a strong non-random association between two types of events, with the strength of association measurable by various kinds of collocation measures. Similar approaches using various collocation measures (also called association or co-occurrence measures), such as t-score, chi-square or log-likelihood, are found in, e.g.,

[1] To the extent the specific meaning is given, this is an onomasiological task (going from meaning to markers). However, we will argue that meanings in a cross-linguistically relevant sense are not extensionally explicit enough at the beginning of an investigation, which is why the task also has a semasiological component (going from markers to meaning).

Cysouw et al. (2007) and Liu et al. (2023). The basic idea is to find forms whose distribution as closely as possibly matches the distribution of interest.

A direct application of collocations, however, fails to account for the necessary distinction between *markers* of a meaning (such as English [eng; Indo-European, Germanic] *not*, *no-*, *never* for negation and *know*(-), *knew* for 'know') and *otherwise associated items* (such as negative polarity items, such as English *but* for negation or the complementizer *that* for 'know'). In other words, while some markers *express* a meaning, others are merely *associated* with it. In order to make the usage-based approach of extracting markers in parallel text corpora applicable to typological studies, it is crucial to find a way to somehow implement in it a distinction that is equivalent to the distinction between markers and otherwise associated items. That is, we are interested in extracting those markers which would in expert judgment be deemed *expressing a meaning*, discarding those markers that are merely secondarily associated with it.

Often markers are more strongly associated with meanings (have higher collocation values) than otherwise associated items, but strength of collocation is not reliable. The algorithm presented in this paper solves this problem by considering the collocation of *sets of markers*, rather than of individual marker candidates.

As an illustration, consider the extraction of forms corresponding to 'know' (modelled with the English lexeme *know*) in French using a rather poor collocation measure, namely Dice, and only word-forms (no character sequences) as candidates. The complementizer *que* 'that' – an otherwise associated item – is the best individual collocating word-form (value 0.21, Dice values range from 1.0 to 0.0, with 1.0 being the maximal value); *connu*, the first intuitively acceptable marker, is only the fifth best individual match. However, if we consecutively assemble marker candidates that improve the collocation value of the entire set and reevaluate already extracted markers for what they contribute to the set, we can obtain quite satisfying results such as [*connu* | *savez* | *connaître* | *connais* | *sais* | *sachant* | *connaît* | *savons* | *connaissent* | *savait* | *sait* | *connue* | *savaient* | *reconnurent*] even with a collocation measure as poor as Dice (set collocation value 0.76). The otherwise associated item *que* 'that' is discarded from the set during the process even though it first appears to be the best individually (see 3.2). We will argue that sets with multiple markers corresponding to a meaning are the rule rather than the exception.

While it is sets of marker candidates that we test for optimal match, individual marker candidates that can make it to the set must at least be associated with the

meaning, so that accidental matches can be excluded. Only recurrent co-occurrence can qualify as association. A quantitatively optimal set of markers could consist of different markers in all contexts of use if they all only occur once. Philologists call forms occurring only once "hapax legomena". However, we do not consider sets of hapax legomena as possible marker-sets for meanings. Put differently, every marker in the set must be recurrent in such a way that it is non-randomly associated with the meaning. This means that a word-form can be included into the set in two ways: (i) either by being a marker itself, and it then has to occur a substantial number of times (for instance, a frequent suppletive marker, such as English *went* for 'go') or (ii) by being a member of the set of word-forms sharing a substring (a morph) that is the non-randomly recurrent marker.[2]

To summarize so far, our approach identifies the markers of a meaning by finding the set of markers which optimally collocates with the distribution representing that meaning; we model both markers and meanings as *sets of discourse contexts* where the marker is attested or the meaning applies. Viewed as sets of discourse contexts, meanings and markers are items of the same kind and hence directly comparable and convertible. A consequence of this choice is that the meaning–marker relationship cannot be considered in abstraction of a particular discourse environment.

A meaning in our approach, then, can basically be any arbitrary set of discourse contexts. However, not all sets of discourse contexts are equally useful and we will argue that to be useful as meaning representations, sets require empirical grounding. As a first step, we can approximate meanings by occurrences of markers in single languages; for instance, the meaning 'know' by where forms of the English verb *know* occur in the English text, but once we have extracted markers for 'know' from many languages, we can determine an "interlingua" (cross-linguistically informed) distribution of 'know' that is not biased to one particular language. Put differently, we can think of meaning as a cross-linguistically comparable concept, as is often done in typology (Haspelmath 2010). However, what we have in mind is a cross-linguistically comparable concept that is not entirely given a priori to the investigation, but that is refined and improved as the cross-linguistic investigation proceeds (see Dahl 2016, who uses the term "generalizing concept").

[2] This is basically Mańczak's (1966: 84) law of differentiation: "More frequently used linguistic elements are generally more differentiated than less frequently used elements" (English translation by Haspelmath 2023: 7).

Our approach, then, is set-theoretical (dealing with collections of objects into sets) in three respects: we operate with (i) sets of discourse contexts expressing different meanings, (ii) sets of discourse contexts reflecting different markers and (iii) sets of markers together expressing meanings. A set of markers expressing a meaning is identified by its optimal collocation with that meaning, which means that there is no other set of marker candidates in that language with a better collocation value. All this can be summarized in (2). We call this suggested mechanism a "law", the underlying idea being that it is generally at work for all sorts of meanings.

(2) *A law of meaning*
A meaning is expressed by the set of non-randomly recurrent markers that together are the best collocation of that meaning

We insist in particular on the word "together". It is the *entire* set of markers that collocates optimally with a meaning rather than its individual markers. As discussed above, the restriction "non-randomly recurrent" is necessary to ensure that each marker in the set is also individually associated with the meaning, which is a much weaker requirement than being best on its own. It is a bit like in football. What matters is not who is the best individual player, but who make up the best team. But even in the best team, every member has to be at least a good football player.

This paper presents a concrete algorithm that implements the law in (2) so that it can be used in roughly sentence-aligned parallel texts.[3] With this algorithm we extract and evaluate the encoding of 'know' and some syntagmatically related domains in a variety sample of 83 languages in digital translations of the New Testament, the only parallel text of considerable length available in a large number of languages from different language families and from all continents. Due to ease of evaluation, the algorithm will first be applied to person names.

The rest of the paper is structured as follows. Section 2 provides background about models of meaning, collocation measures and the four domains investigated in this study (negation, 'know', first person subject and propositional complementizers). Section 3 demonstrates how the law can be implemented into a practical algorithm and illustrates how the algorithm works. The language sample is introduced in 3.4. Section 4 presents results and analysis for the four domains surveyed. The discussion

[3] Actually, Bible verses rather than sentences are used, and these often contain several sentences.

in Section 6 puts the results obtained into a larger context and section 7 concludes the paper.

2. Background

2.1. Coexpression with and without implying semantic atoms and cross-linguistic equivalence

Usage-based massive cross-linguistic comparison has revealed considerable cross-linguistic semantic diversity, which is often approached with semantic maps modelling semantic space (see Georgakopoulos & Polis 2018 for an overview). According to François (2008), the semantic map approach allows us to break up "polysemous lexemes of various languages into their semantic 'atoms' or senses", which can be arranged in "an etic grid against which cross-linguistic comparison can be undertaken" and "[l]anguages differ as to which senses they colexify, i.e., lexify identically" (François 2008). Since grammar does not differ much from the lexicon in this respect, the notion of colexification has been generalized to coexpression, "the availability of two meanings for a minimal form in different contexts" (Haspelmath 2023: 1; Hartmann et al. 2014). However, in practice, the phenomenon of coexpression does not presuppose semantic atoms (primitive semantic units), but can be applied to any sort of analytical primitives (Wälchli & Cysouw 2012: 679). It is therefore problematic to view coexpression as deviation "from the canonical ideal of a one-to-one correspondence between meanings and shapes" (Haspelmath 2023: 2). Usage-based typology, especially distributional approaches and notably the study of massively parallel texts, has revealed that finding categories that are extensionally fully equivalent across two languages is rare if data is not sparse. One solution is to approximate senses bottom-up by clustering (see, e.g., Beekhuizen et al. 2023: 443, 445 and the literature mentioned there). However, while not affecting the practical usefulness of the notion of coexpression, findings from typological corpus-studies strongly question whether the canonical ideal of a one-to-one correspondence between one meaning and one marker is of any use as it fosters categorial

particularism ("descriptive formal categories cannot be equated across languages"; Haspelmath 2010: 663).[4]

Aside from proposing a practical solution for how the markers expressing a meaning can be identified in parallel texts, this article has a theoretical aim, which is to argue that the idea of a canonical ideal of a one-to-one correspondence between one meaning and one marker (or shape or form or construction) that pervades linguistic approaches of most different kinds is mistaken. We will show that abandoning it does not necessarily result in "rampant many-to-many relationships" that only obscure matters (Haspelmath 2010: 680), but in meaning–marker relationships that can still be uniquely determined. The law formulated in this paper is a suggestion for how meaning–marker relationships can still be uniquely determined for all kinds of meanings, both lexical and grammatical. The question as to whether categories can be equated across languages then boils down to what we mean by "equate". If we mean "complete identity in extension" and "one-to-one relationship" (exact congruence), we agree with Haspelmath (2010) that the answer is "No". But if we mean uniquely determinable relationship following a general principle, our answer is "Yes".

2.2. Elucidating the law

In this section, the main three 'ingredients' of the proposed law are discussed – meaning, sets of non-randomly recurrent markers and best collocation. First, our approach to meaning is presented and compared with more traditional views. We also briefly compare and contrast our view with those taken in set-theoretical formal semantics, compositional semantics, Natural Semantic Metalanguage and construction grammar. The relationship of our view on meaning and the popular notion of colexification or coexpression is also further developed. Secondly, what we mean by sets of markers is discussed. Markers are compared to notions such as lexeme and morpheme and our view of marker sets as coalition phenomena is outlined. Thirdly, the notion of collocation and how to measure it is discussed. We distinguish

[4] Interestingly, the ideal of one-to-one correspondence is retained also in NLP-approaches to parallel texts dealing with colexification: "We define crosslingual stability of a concept as the degree to which it has 1-1 correspondences across languages, and show that concreteness predicts stability" (Liu et al. 2023).

between inter- and intra-text collocation, and present a number of collocation measures, including the Dunning log-likelihood which is used in this paper.

2.2.1. Meaning

Our approach to meaning is *discourse*-based. We argue that the meaning–marker relationship cannot be properly studied in what corresponds to *langue* or competence in models such as de Saussure's or Chomsky's. However, our approach differs from most discourse-oriented approaches in targeting primarily language structure below rather than above the levels of sentence and clause and by considering discourse phenomena stochastically rather than as individual events. Thus, meaning in this article is conceived of

> (i) distributionally (as a property shared by a set of discourse environments) rather than exemplar meaning and
> (ii) extensionally rather than intensionally.

In practice, this means that we conceive of a meaning as a *set of discourse contexts*.[5] As such, the approach is *set-theoretical* and may be faintly reminiscent of set-theoretical approaches in formal semantics, although it is actually quite different. Formal semantic approaches, such as Montague Grammar, target referents and truth values by means of sets by assigning sets of individuals in real and possible worlds and set of truth values to semantic values (see, e.g., Dowty 1979).

Our approach does not address the relationship between markers and referents, which is a major concern of formal semantics, but it is certainly compatible with formal semantic approaches, although this is not developed in this article.[6] Note also that there may be an analogy to possible world semantics where possible (more or less probable) discourse occurrences are involved.

---

[5] We do not claim that meaning *is* linguistic usage distribution. We only claim that there must be currency conversion from meaning to markers and from markers to meaning, which is why meaning must have some sort of manifestation where it has the same properties as markers, and this is extension in usage. Hence, our approach is compatible with any theory of meaning that contains a component where meaning manifests as actual and fully explicit linguistic usage distribution.

[6] It is probably more profitable to model sets of referents on the basis of marker tokens than to model sets of referents directly from marker types.

This article only considers attested sets of occurrences in corpora, but the model could be further expanded probabilistically to include even non-attested and future discourse environments (see Table 1).

Well-known *intension-based* models of the meaning–marker relationship are found in de Saussure's structuralism and Crofts Radical Construction Grammar, relying on symbolic links between marker and meaning, as illustrated in Figure 1.

	Attested	**Not attested**
Sets of referents...	...in the real world	...in possible worlds
Sets of occurrences in discourse…	...in existing accessible corpora	...in possible (future or not attested) discourse environments

Table 1: Analogies between two different set-theoretical approaches to semantics.

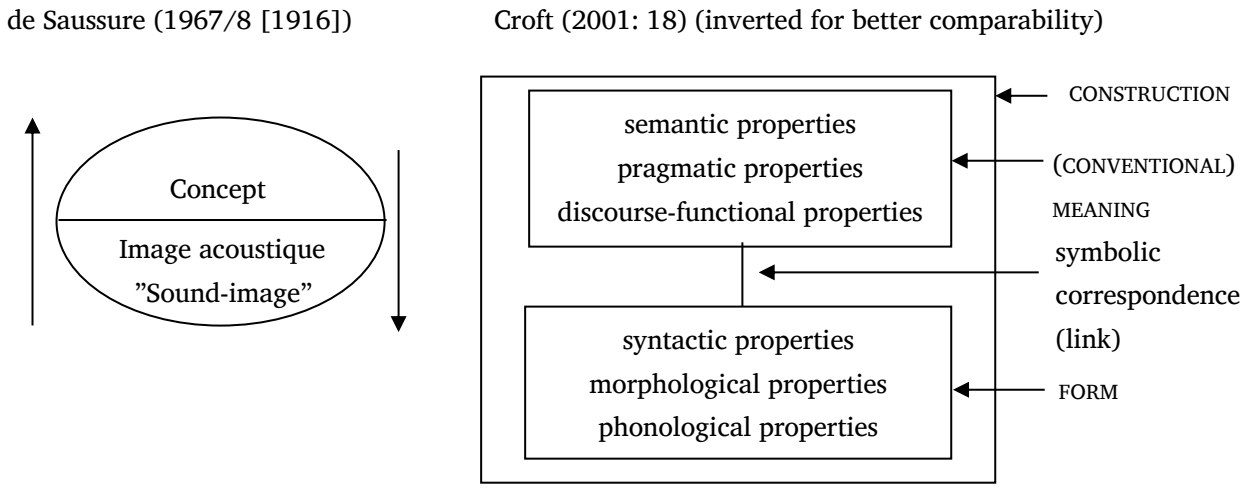


Figure 1: Symbolic link between marker and meaning in structuralism and in construction grammar.

This paper offers an alternative by modelling the meaning–marker relationship by way of *extension* – i.e. as sets of discourse occurrences – as shown in Figure 2. Meaning is linked extensionally, via optimal match, to a marker set.

An advantage of the extensional approach is that it can deal with different models of meanings. Semanticists do not agree whether meaning is strict (a so-called Aristotelian definition) or fuzzy (core prototypical vs. peripheral less-prototypical exemplars) and with an extensional approach we need not decide. Distributions can be modelled both as strict and as fuzzy sets.

A consequence of this approach is that any set of contexts can be used as a meaning.

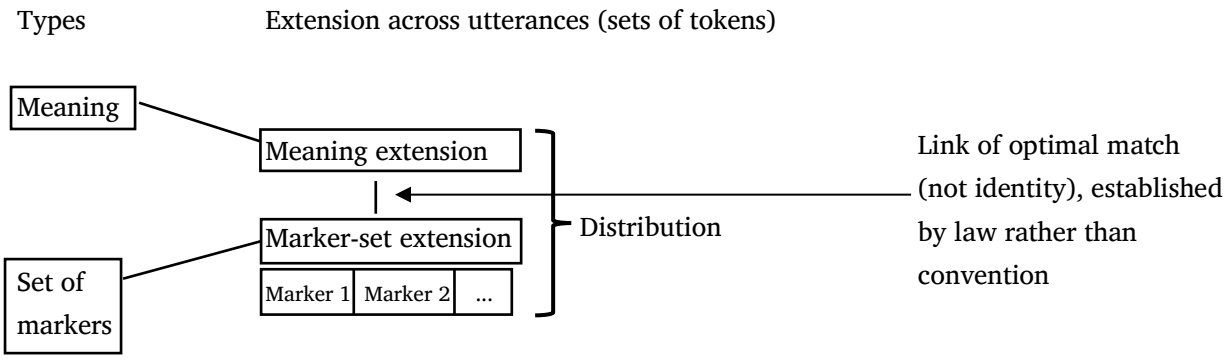


Figure 2: Marker and meaning are linked via extension.

However, not every set will be empirically well-founded or work well. We may be inclined to postulate criteria for which sets of discourse contexts qualify as profitable models of meanings in terms of intensional semantic criteria (whether the tokens have at least family resemblance) or extensional semantic criteria (whether they cluster to a region of semantic space). Different sets of discourse contexts modelling a meaning can be evaluated within the method proposed here by how well they function as the basis for cross-linguistic investigations in terms of resulting *coverage* (proportion of contexts for which a marker has been found) and *dedication* (proportion of all contexts using a marker from the set that are located within the search domain).

We distinguish *parochially expressed meanings* and *interlingua meanings*. A meaning that optimally fits a set of language-specific markers can be said to be a parochially expressed meaning. For instance, the set of occurrences of the English lexeme *know* or the pronominal form *I* are parochially expressed meanings. Interlingua meanings minimize language-specific bias in cross-linguistic investigations and reflect the aggregated patterns of distribution of forms in many different languages (a more detailed description of how these are arrived at is given in section 4.2). We expect that interlingua meanings will work better.

Similarly, it may also be considered what the optimal level of generality for matching meanings with markers is (illustrated in Table 2). We do not believe that it is possible to arrive at semantic atoms when picking subordinate-level concepts. Our hypothesis is that it will be *basic-level concepts* that are easiest to match directly with marker sets, which is well in line with prototype theory and other approaches in cognitive linguistics (Rosch et al. 1976) and also with Natural Semantic Metalanguage. Subordinate-level concepts are often lexicalized in particular

languages, but this can be accounted for by not expecting markers and meanings to match one-to-one. Basic-level concepts have higher text frequency than subordinate-level concepts, which makes them easier to approach stochastically. Superordinate-level concepts, however, may be too frequent in texts to allow for clear distributional patterning and their sets of markers can be too large.

Subordinate-level concept	**Basic-level concept**	**Superordinate-level concept**
'know a person personally', 'know a person', 'know a fact'	'know'	cognition predicate, cognition/perception predicate, experiencer predicate

Table 2: Level of generality of meanings.

Different levels of concepts may also be applied to account for our approach's take on the popular notions of colexification or, more generally, coexpression. What may be viewed as a marker-set–meaning relationship (such as (3) for 'know' in French) may at another level be viewed as a case of colexification (*know* corresponding to French *savoir* for 'know (that)' and *know* corresponding to French *connaître* 'know person') or as a case of dislexification (*savoir* for 'know that' and *connaître* for 'know person').

(3) Marker-set for 'know' in French
[#sav | #conn | #sach | #sais# | #sait | #saur | #su#] (# stands for word-boundary)

Since meanings are not fixed primitive units but sets of discourse occurrences and the marker–meaning relationship is conceived of not as an intensional one-to-one link but as a case of best match, many levels of analysis are available simultaneously. Choosing a higher level of abstraction risks concealing patterns that are of explanatory interest in certain languages (such as the *savoir/connaître* distinction), whereas choosing a lower level may conceal more general cross-linguistic patterns (such as that most languages do in fact colexify 'know that' and 'know person', see Sjöberg 2023).

For the purposes of demonstration in this paper, we have chosen meanings which we intuitively believe to be intensionally plausible and designed the distribution in such as a way as to make them extensionally plausible. However, there is nothing saying that the meanings chosen cannot profitably be modelled as either containing some sub-meanings or being part of super-meanings and the results of this be

expressed in terms of coexpression. Our approach allows for this, and in fact allows for a quantification of the appropriateness of a given level of concept for a certain language or in aggregate.

To put things in terms of *semantic features* or *semantic decomposition*, our approach operates with a single semantic feature (the meaning searched for) and decomposition takes place in a binary way (the meaning searched for against anything else).[7] As a consequence, no distinction is made between simple meanings (only one feature or semantic prime) and complex meanings (a combination of several semantic features, recently termed "synexpression" by Haspelmath 2023: 1, "the simultaneous presence of two meanings in a minimal form"). In other words, the present approach treats all meanings the same way: as *one-feature non-decompositional meanings*. In this article, this is illustrated with the meaning first person singular subject which is usually considered a complex meaning with an arguably lexical component (first person singular) and an arguably grammatical and syntactic component (subject). We will show that our approach can handle both traditionally simple and complex meanings. In either case, the meaning can be modelled extensionally as a set of discourse contexts that can be matched to a set of markers directly.[8]

Among decompositional approaches, there is one that is of great theoretical and practical interest to us, even though we do not share its decompositional stance – the framework of *Natural Semantic Metalanguage* (NSM), "a decompositional system of meaning representation based on empirically established universal semantic primes" (Goddard 2008: 1), originally developed by Anna Wierzbicka. The reason is that unlike most other decompositional approaches, NSM does not postulate abstract semantic features, but operates with non-decomposable (that is, primitive) lexical concepts assumed to be expressed by markers in all languages. NSM-work is very important for us due to its interest in identifying markers in all languages for a number of very general lexical concepts, some of which can be said to be intermediate between lexicon and grammar (such as 'I' and 'not'). Unlike NSM, we do not assume that there is any privileged set of "primes". Rather, many more meanings than those considered to be primes in NSM can be expected to be universally or almost

[7] This is called "local decomposition" in Wälchli & Sölling (2013: 86).

[8] In terms of a qualitative approach, first person singular subject stands in relation of a hyponym to first person singular, a statement which can be made without adducing to the notion of complex meanings. Hyponym (without any connotation of taxonomy) aligns better than "synexpression" with the set-theoretical approach pursued here.

universally marked across the languages of the world (for 'only', see Wälchli 2024). However, we share NSM's interest in finding meanings that are "expressible by words, phrases or affixes in all or most of the world's languages" (Goddard 2012: 718) and three of the four concepts considered in the main part of this article, negation, 'know' and 'I', happen to be postulated as semantic primes in NSM. However, we do not expect our approach to work for NSM-prime-concepts only; we apply the same method also to proper names, which NSM has notorious difficulties in accounting for. NSM accounts for lack of one-to-one relationships between meanings of primes and markers among other things by polysemy (Goddard 2008: 5), for which we use the more comprehensive notion coexpression. NSM is also of interest to us, because – unlike most modern cross-linguistic approaches – it focuses on markers rather than constructions.

Let us now turn our attention to *construction grammar* (see, e.g., Goldberg & Suttle 2010 for an overview), with which the current approach shares many features, notably its usage-based design, the lack of a strict distinction between semantics and pragmatics and the high importance assigned to item-specific information. Can a meaning and a set of markers expressing it be considered a *construction*? According to constructionists, "language consist of systematic collections of form-function pairings, or *constructions*" (Goldberg & Suttle 2010: 468). Constructionists emphasize that forms need not be minimal units and can be segmentable wholes, which is in complete accordance with our approach. Markers need not be morphemes, but can consist of a sequence of several markers or even word-forms. That such non-minimal items systematically entertain item-specific relationships to meanings is a core contribution of construction grammar theory. We could say that constructionists emphasize the *syntagmatic non-compositionality* of languages; an item that can be considered a set of smaller units co-occurring in a construction can have a meaning of its own. For instance, in Yélî Dnye (Yele; yle; Isolate, New Guinea), 'know' is expressed by the possessive pronoun together with *lama* 'knowledge' and not by a single morpheme. However, what our approach emphasizes in addition is that sets of forms occurring at different places in discourse also can entertain direct links to meaning; we may call this *paradigmatic non-compositionality*. For instance, 'know' in French is not expressed by a single marker, but by a set of markers, such as [*#sav* | *#conn* | *#sach* | *#sais#* | *#sait* | *#saur* | *#su#*], occurring in different contexts. Just as co-occurring units together as a whole may be said to be linked to a single meaning, we argue that such a set also, as a whole, can be viewed as linked to a single meaning. Our impression is

that most approaches to construction grammar operate on the basis of an ideal of a one-to-one relationship between meaning and marker which is not compatible with our approach, but this does not seem to be an intrinsic requirement of the constructionist approach.

There does not seem to be any fundamental contrast between segmental markers and constructions. Many construction types, such as n-grams, including hybrid n-grams (Wible & Tsao 2010), and pivot schemas (e.g. *more* _ in *more milk, more grapes, more juice*; Tomasello 2003: 114), feature segmental markers. However, further issues concerning constructions are of practical rather than fundamentally theoretical nature. The concrete implementation of our approach implies that items that can be included in marker-sets must be accessible among limited sets of possible candidates. The more abstract a construction, the more difficult it is to conceive of it as a member of an enumerable type of marker candidates. This is why candidates in the present article will be limited to word-forms, morphs and bigrams. Put differently, abstract constructions are a considerable practical challenge for us. But abstract constructions are not excluded from our method as long as there are ways to access them by starting from accessible limited sets of candidates.

2.2.2. Sets of markers

Markers are a central ingredient in our approach. These are neither lexemes (or gramemes) nor morphemes. Haspelmath & Sims (2010: 333) define *lexeme* as "a word in an abstract sense; an abstract concept representing the core meaning shared by a set of closely-related word-forms ... that form a paradigm" and *grameme* might be defined in analogy as a grammatical marker or construction in an abstract sense with a meaning shared by a set of closely-related grammatical morphemes or constructions. Since our approach identifies sets of markers expressing the same meaning, there is no need to deal with lexemes or gramemes separately.

Lexical and grammatical meanings are very commonly expressed by several different markers, a phenomenon termed *polymorphy* in Wälchli (2014: 359).[9] One reason for this is that cumulative expression of several different recurrent meanings by a single marker allows for high density of information in discourse – for instance

[9] NSM uses the term allolexy for "a situation in which there are multiple lexical realisations of a single prime" (Goddard 2008: 6).

combining the present and person marking in a single morpheme. The resulting set-character of markers corresponding to a meaning – only the set of combined tense-person markings can be said to express the present tense – entails that sets of markers, such as lexemes and grammatical categories, are *coalitions*. Like in democratic elections without clear majorities, a single party cannot form a government – a coalition is needed. Different markers join forces opportunistically (because this is what the environment requires them to do) in order to be able to optimally match a meaning. From this perspective, lexemes are not necessarily basic or fundamental notions of analysis. Lexemes are nothing else but a special kind of opportunistic coalition of markers. It is thus neither theoretically necessary nor practically particularly useful to group markers systematically to lexemes or gramemes before linking them to lexical or grammatical meaning.[10]

A *morpheme*, "the smallest meaningful part of a linguistic expression that can be identified by segmentation" (Haspelmath & Sims 2010: 335), can be determined only after all meanings at work in an expression have been considered, whereas our approach only considers one single meaning at a time. There is therefore no direct relationship between markers and morphemes in our approach. A marker can be a single morpheme (or rather an allomorph, if a morpheme has allomorphs), a sequence of several morphemes, a word-form or two subsequent word-forms or whatever the definition of marker candidates allows for – the method simply does not take the notion of morpheme into account. In this respect, our approach is similar to construction grammar, where meaning is not necessarily paired with the smallest parts in form (see 2.2.1). If we consider just one meaning at a time, the part-whole relationship simply does not apply.

It is important to note that most research in morphological theory is heavily influenced by structuralism (considering all markers and meanings in their interplay in a system), whereas our approach is *anti-structuralist* in considering only one meaning at a time. Meanings are not considered in their interplay in the system, but in isolation.

A further important point is that markers are very different from *citation forms*. One of the advantages of our approach is that we can entirely do away with citation forms, which are not only language-dependent, but even grammarian-dependent, and which

[10] Operating with word-forms instead of lexemes also has practical advantages for low resource languages where lemmatizers are not available (Schütze & Asgari 2017: 115).

are an obstacle for cross-linguistic comparison of markers. Our approach contributes cross-linguistically directly comparable markers, since they are determined in exactly the same way for all languages addressed. However, what can be extracted as a marker is strictly constrained by what kinds of marker candidates we allow for. It is therefore essential that considering what markers there are goes hand-in-hand with the study of what kind of marker candidates there can be.

Finally, it is unfortunate that our practical application is entirely dependent on written form, which induces a heavy written-language bias. If phonological input was available, this could be avoided.

2.2.3. Collocation

Going from meaning to form (onomasiology), we have to start with some sort of search distribution modelling a meaning. Since we cannot expect that the search distribution is completely identical with the target distribution, but only similar, we need a way to assess what it means to be sufficiently similar to establish a meaning–marker link. This can be done by means of measuring the strength of collocation. Addressing meaning by way of collocation is in the spirit of Firth's (1957) famous saying "you shall know the word by the company it keeps" (1957: 11). Firth in his turn refers to Wittgenstein's (1958) famous saying "the meaning of words lies in their use".

A collocation is traditionally defined as "an expression consisting of two or more words that correspond to some conventional way of saying things" (Manning & Schütze 1999: 151) and corpus linguists often use collocations to show subtle differences between near-synonyms, such as *strong* and *powerful*, which differ in collocates (for instance, *strong tea*, but *powerful drugs*). Whereas in monolingual corpora collocations are used to investigate which words go together, in parallel corpora, we can investigate how forms go together with their translation equivalents (Cysouw et al. 2007; Dahl 2007) and translation-equivalents can be used to model meaning. The major difference is whether the collocates are overtly present in the text.

In both cases, the basic idea is to compare the occurrence of some entities and to determine whether the presence of one reliably informs on the presence of the other. This may be done for words within a text, as in the example of *strong tea*. Given the presence of *tea*, we have reason to expect the presence of *strong* (at least in comparison

to other adjectives). This might be referred to as *intra-text collocation*. However, given some way of matching the place of occurrence across texts – as parallel corpora give – we can also consider what might be referred to as *inter-text collocation* (called trans-co-occurrence by Cysouw et al. 2007: 159). We are not considering here whether the presence of one marker predicts the presence of another marker *within* the same text, but whether the presence of a marker in one text predicts the presence of a marker in another, parallel, text. Now, if we think of the marker in another language as being similar to a cross-linguistically generally applicable meaning modelled as a set of occurrences, inter-text collocation of markers is very similar to *meaning–marker collocation*, which is what we are interested in in this article. Put differently, we use inter-text collocations in a parallel text corpus to model meaning–marker collocation. This is all summarized in Table 3.

Type of collocation	Examples
Intra-text collocation of markers (in a corpus)	*Strong* collocates with *tea* but not with *powerful*
Inter-text (intra-language) collocation of markers (in a parallel text corpus)	Forms of the English lexeme *know* collocate with French word-forms such as *connu, savez, connaître* etc.
Meaning–marker collocation (in a parallel text corpus)	The semantic comparative concept 'know' collocates with French word-forms such as *connu, savez, connaître* etc.

Table 3: Three types of collocations.

Given the *optimality-based* nature of this approach, it is faintly reminiscent of Optimality Theory (OT), a linguistic theory according to which surface forms of a language result from optimally satisfying conflicting constraints (see, for instance, McCarthy 2007). Like OT, our approach operates with candidates, but these are not generated by the model, but are given as types of surface strings (such as word-forms). Our approach does not work with constraints.

The literature reports a considerable number of collocation measures, some of which are practically illustrated in Table 4 with the best word-form and bigram collocations from the French Darby NT translation matching the lemma 'know' in the English American Standard NT translation. For a survey, see Manning & Schütze (1999: 162-176). It can be seen that especially less sophisticated collocation measures, such as Dice and *t*-score, do not distinguish between markers (here forms

of *savoir* ‘know that’ and *connaître* ‘know person, thing’; in boldface) and otherwise associated items of ‘know’, such as complementizer (*que*), negation (*ne*, *pas*) and first and second person pronouns (*je*, *vous*).

Dice			**t-score**			**LogL (Biemann et al. 2004)**		
1	0.2106	w que	1	8.2605	w que	1	9.7271	w **connu**
2	0.1926	w ne	2	6.7849	w ne	2	9.3354	w **sais**
3	0.1702	w vous	3	6.5075	w **connu**	3	8.8402	w **savez**
4	0.1682	w je	4	6.4209	w **sais**	4	8.7462	w **savons**
5	0.1664	w **connu**	5	6.2245	w **savez**	5	8.2109	w **sachant**
6	0.1627	w **sais**	6	6.154	w **sachant**	6	7.261	w **connais**
7	0.1621	w pas	7	6.1142	w **savons**	7	7.1656	b nous **savons**
8	0.1536	w **savez**	8	5.5832	w **connais**	8	6.2722	b vous **savez**
9	0.1518	w **sachant**	9	5.5162	b nous **savons**	9	6.1651	w que
Cosine			**Phi**			**Dunning’s LogL**		
1	0.3	w **connu**	1	0.0426	w **connais**	1	453.90	w **savons**
2	0.29	w **sais**	2	0.0423	b vous **savez**	2	426.95	w **connu**
3	0.28	w **savons**	3	0.0417	b **sachant** que	3	368.20	w **sais**
4	0.28	w **savez**	4	0.0413	w **connaissez**	4	358.18	w **savez**
5	0.27	w que	5	0.0410	b **sais** que	5	352.57	b nous **savons**
6	0.26	w **sachant**	6	0.0403	b **ne sais**	6	316.65	w **connais**
7	0.26	w **connais**	7	0.0401	b **connais** pas	7	279.18	b **savons** que
8	0.26	b nous **savons**	8	0.0039	b vous **connaissez**	8	270.09	w **sachant**
9	0.24	b vous **savez**	9	0.0389	b de **connaître**	9	261.91	b vous **savez**

Table 4: The best word-form (w) and bigram (b) collocates in French (Darby) for lemmatized English ‘know’ (American Standard).

Collocation measures are computed on the basis of values such as the following:

A: Number of occurrences in the given distribution,
B: Number of occurrences in the test distribution,
A∩B: Number of occurrences shared by the given and the test distribution, and
N: Total number of occurrences

In our application to the New Testament, number of occurrences can simply mean number of verses, so N is the number of verses of the New Testament (which may slightly vary from translation to translation, so we take the number of verses in a version of Koine Greek as basis).

In this paper, we will use Dunning's log-likelihood (Dunning 1993; see also Appendix I)[11], as it has a number of advantages. The log-likelihood ratio test is more appropriate for sparse data. The test value $-2\log\lambda$ is asymptotically χ^2-distributed if the expected values in the 2-by-2 contingency table are not less than 1.0 (Manning & Schütze 1999: 174). The threshold can thus be aligned with a confidence level (for 0.005, the threshold is 7.88). This lower limit for the threshold assures that extracted forms are at least in some way non-accidental. However, otherwise associated items are as non-accidental as markers, and semantically related concepts (such as co-hyponyms or antonyms) are often also associated in texts. Texts can also contain repetitions that blur the picture. This is why, we will have to use higher thresholds, somewhere in the range between 20 and 210. The level where undesirable corpus-specific collocations start occurring differ from meaning to meaning and it is therefore useful to set thresholds individually for each meaning after manual evaluation.[12] For instance, the meaning 'bird(s)' (used and exemplified in Liu et al. 2023) often occurs in the same verses as 'reptiles', which is why our method with alignment by verses requires a rather high threshold (around 61) for 'bird(s)' in order to avoid forms for 'reptiles' being extracted.

2.3. The four meanings to be considered

In Section 4, we will consider four different lexical and grammatical meanings: negation, 'know', first person subject ('I') and propositional complementation ('that'), all being frequent in language use. In 2.2.3 we have seen that negation, first person subject and complementizers are otherwise associated items of 'know', so there is a considerable overlap in occurrence, which is a major motivation for considering exactly these four meanings together in this article. Examples (4) and (5) both instantiate and illustrate three of the four meanings at a time.

(4) know, first person subject and complementation (eng-x-bible-lexham, 43008037)
I know that *you are descendants of Abraham*

[11] Appendices, including information concerning the corpus (translations of the New Testament) are available at https://doi.org/10.5281/zenodo.10522345.

[12] See, for instance, Beekhuisen et al. (2023: 438) for emphasizing that evaluation should assure a reliable quality of extraction.

(5) know, first person subject and negation (eng-x-bible-lexham, 42022057)
***I** do **not know** him!*

This overlap in use entails that we can expect a certain amount of overlap in encoding in some languages. For instance, suppletive forms for 'not know' are expectable results both for the meaning 'know' and for the meaning negation, which illustrates that no meaning has exclusive rights to any marker. A marker can be part of several marker sets, expressing several meanings, at the same time. However, the four meanings can also be taken to be illustrative of a general law since they are all very different meanings, with 'know' being the most lexical one and complementation the most grammatical one. Negation and first person are often considered grammatical meanings, but they also figure in the list of semantic primes or "universally lexicalised meanings" in Natural Semantic Metalanguage (NSM) (Goddard 2008: 5). When addressing first person, we will actually look at the meaning first person singular subject, which is a combination of the lexical meaning first person singular and the grammatical relation subject, in order to illustrate that the approach can be applied to lexical meanings and to grammatical meanings and to mixtures of lexical and grammatical meanings alike.

Negation is one of the best investigated domains in typology. However, most studies concentrate on certain subdomains of negation. Miestamo (2005) focuses on standard negation, "the basic way(s) a language has for negating declarative verbal main clauses" (2005: 1), which excludes, for instance, prohibitive (negative imperative), existential negation, non-finite negation and negative indefinite pronouns (these and other subfields of negation have been studied in separate typological investigations). Restricting typological studies to standard negation or prohibitives or negative indefinite pronouns is very useful if the mechanisms of negative constructions are to be considered. However, here we take a more holistic approach and want to consider how negation is marked in general, glossing over the many subtleties of constructions of negation. Following from our focus on distinguishing true markers from otherwise associated items, a very important distinction for us is the one between negation markers and negative polarity items. The distinction cuts across such domains as negative indefinite pronouns. Haspelmath (2013a) distinguishes negative indefinite pronouns that always co-occur with predicate negation, such as Afrikaans (afr; Indo-European, Germanic) *Wanneer jy mense help, mag **niemand** daarvan weet **nie*** 'When you help people, no one should know about it' (afr-x-bible-boodskap, NT 40006003),

from negative indefinite pronouns that never co-occur with predicate negation, such as English ***no one*** *should know about it* and languages with mixed behavior. In Afrikaans, *nie* is the negation marker and *niemand* is just a negative polarity item. In English, however, *no (one)* is a negation marker. Put differently, if the algorithm extracts *niemand* for negation in Afrikaans, this is a mistake, but if the algorithm misses *no* in English, the English negation marker set is incomplete. The well-established distinction between negation markers and negative polarity items makes negation a very useful test domain for evaluating our approach.

Despite well-known connections to perception verbs (Sweetser 1990; Evans & Wilkins 2000), the ***'know'*** domain is cross-linguistically quite distinct from perception and from other cognition domains (Sjöberg 2023). This makes 'know' a good test domain for our purposes. We also chose it notably because Sjöberg (2023) contains a typological investigation of 'know' in 83 languages based on data from the NT and we can use this sample for evaluation. Sjöberg (2023) shows that there is a great deal of internal lexical variability in the 'know' domain. For instance, many languages distinguish between 'know (person)' and 'know (fact)' and many languages have lexical negative 'know' verbs ('be ignorant'). Knowledge verbs can also be quite irregular (the same lexeme has several rather different stems and forms, such as French *sav-*, *sach-*, *sait*, *su*). Whether all languages have 'know' expressions is a matter of discussion. In Natural Semantic Metalanguage, 'know' is considered a semantic prime (Wierzbicka 2018). However, Pawley (1994) has argued that Kalam (kmh; Nuclear Trans New Guinea, Madang) lacks 'know' since there is only a very general perception and cognition verb *nŋ-* <niŋ->, for which Pawley & Bulmer (2011: 416) list twenty-two translation equivalents including 'be conscious; be awake; think; know; perceive; see; look at; hear; listen; feel; smell; taste; try; learn; be used to; believe'. Pawley (1994: 394) emphasizes that the verb stem *nŋ-* <niŋ-> alone stands for 'know' in Kalam and that there is no other element that expresses 'know' together with *nŋ-* <niŋ-> in a construction. Kalam happens to be a language in our sample, so we can test whether <niŋ-> is extracted.

English *I* is *first person subject* (conflating intransitive subject S and transitive subject A) and the Anglocentric and Eurocentric notion of subject as a fundamental grammatical relation in syntax has received highly privileged treatment in most syntactic theories. Here we treat it distributionally and semantically exactly like any other meaning, which may provide a complementary perspective to syntactic approaches, such as surveyed in Haspelmath (2013b), who distinguishes between

pronouns (free person forms having the same syntactic function as noun phrases) and indexes (bound person marking on verbs, auxiliaries and as clitics). In many languages, person marking can be expressed both by pronouns and indexes and the question arises as to whether we should simply view such multiple marking as double exponence (Haspelmath's "double expression view") or whether either pronouns or indexes should be considered the sole argument (either pronoun arguments with agreement or bound arguments with pronominal appositions or adjuncts, as Jelinek 1984 has suggested for Warlpiri [wbp; Pama-Nyungan, Desert Nyungic] in a classical article).

The final task addressed in Section 5 is to retrieve *complementizers* such as English *that* and related markers from the languages of the sample. Complement clauses, traditionally understood as subordinate clauses having the function of an argument (with main verbs such as 'see', 'hear', 'know', 'believe', 'think', 'say' and 'want'; Dixon 2006; Noonan 2007), are a typical instance of a syntactically a priori defined category type. There is much reason to believe that complement clauses are not prototypical subordinate clauses since what is commonly considered the main clause often functions as an epistemic marker, a marker of illocutionary force or is just a parenthetical (Diessel & Tomasello 2008). We will focus here on contexts where non-controlled, embedded, declarative, propositional, factive and finite clauses are most expectable. This excludes, for instance, direct speech, indirect questions and state-of-affairs complements, such as 'how to play the piano' (Kehayov & Boye 2016: 3), and happens to favor cognition rather than perception, the latter being more inclined to be expressed with some sort of non-finite construction (Horie 1993). As we will see, a major challenge in extracting markers of this kind of clauses is that 'know' is so strongly represented, at least in the NT, that it is difficult to avoid 'know'-markers in the extraction. Once this problem is addressed, complementation turns out to behave quite similarly to the other meanings treated in this paper.

3. Method and data

3.1. Introduction

In this section, we will first demonstrate how the law formulated in (2) can be turned into an algorithm that we implement in a Python program (3.2). Section 3.3 deals with otherwise associated items and how they are expected to relate to the algorithm.

We will then introduce the sample of 83 languages (3.4) to which we apply the algorithm in this paper. Finally, we will illustrate how the algorithm works with the easiest task there is: to find proper names (3.5). For a comparison of our method to earlier approaches in the literature, see Appendix A.

3.2. Turning the law into an applicable algorithm

The law as formulated in (2) does not say anything about how the optimal set of markers for a meaning can be found. Given that there are very many candidates that all might be included or not included into the set in all sorts of combinations, the task of finding the best set is not entirely trivial. For making the law practically applicable for cross-linguistic comparison, we will confine ourselves (i) to searching for a *semi-transparent* set of markers (rather than for an entirely opaque set) (ii) by applying *one uniform search procedure* (rather than a whole battery of different alternative search procedures) and (iii) to *directly accessible candidate sets* (rather than opaque candidate sets).

(i) A semi-transparent set implies that it must always be clear how to decide on the next step to take (including the first step, the first candidate to be selected). This entails that at least one marker (the first one to be selected) must have high cue-validity. Hereby we exclude solutions that are entirely opaque and can be found only by trial-and-error. (ii) In linguistic typology, it is important to compare like with like. Algorithmically, this means applying exactly the same search procedure to all languages considered. It is therefore preferable to have a uniform search procedure for finding the optimal set of markers that is applicable to all languages. (iii) In unannotated texts, search strings such as *word-forms* (character sequence between two spaces), *morphs* (continuous character strings within word-forms) and *bigrams* (sequences of two-adjacent word-forms) are directly accessible types of marker candidates (see Table 5).[13] This excludes discontinuous markers including all kinds of non-concatenative morphology, which is a provisional solution. Let us simply see how far we can get with very simple sets of marker candidates only.

[13] For practical reasons, we ignore the difference between orthography and phonology. It would be better to have all texts in phonological notation, but we have to go for what is available. The simple types of marker candidates that we choose have the advantage that they are not particularly sensitive to phonology.

(i)	All *word-form* types in the text (whatever string is separated by space), e.g. *knowing*
(ii)	All potential *morphs*; that is, all continuous sequences of characters within word-forms, e.g. *#kn*
(iii)	*Bigrams*; that is, all sequences of two word-forms in running text, e.g. *knowing that*

Table 5: Three sets of candidates.

The choice of transparent candidate types implies that markers are not lexemes, but just recurrent strings. No lemmatization is applied. There are no such things as citation forms in our approach. Thus, a formally variable verb such as French *savoir* 'know' will not be represented by a single arbitrarily chosen citation form as the infinitive, but rather by a set of characteristic strings, some of which are stems, such as *sav-* and *sach-*, and some of which are salient word-forms, such as *sait* and *sais*.

The algorithm applied (for pseudo-code, see Appendix B) has the following ingredients and properties:

(a) *Candidates*: It is applied to directly accessible candidate sets: word-forms (w), morphs (m) and bigrams (b).

(b) *Ranking order*: Candidates are considered for selection in a ranking order determined by their individual collocation value with the search distribution (the meaning to be expressed). Dunning's log-likelihood is used as collocation measure. See Table 6 for an example.

(c) *Selection*: Going through the entire set of candidates in ascending ranking order, a candidate is selected (provisionally included into the set) if the set containing it has a collocation value that exceeds the collocation value of the set lacking that candidate by at least the threshold. This means that the highest ranked candidate, which is the first one considered for selection, is always selected if its collocation value exceeds the threshold. The same collocation measure, Dunning's log-likelihood, is used. See Table 7 for an example.

(d) *Reevaluation*: Once all candidates have been considered for set inclusion, all selected candidates are reevaluated. A candidate is removed from the set if the collocation value of the set including it does not exceed the collocation value of the set lacking it by at least the threshold. This is, among other things, a possibility to remove the candidate with the best individual collocation value if it does not contribute to the optimality of the entire set of markers. The same collocation measure, Dunning's log-likelihood, is used. Reevaluation often does not change anything; in the examples in Tables 7 and 8 it has no effect.

(e) *Output*: Extracted markers are ordered according to how much they contribute to the set as measured in Reevaluation. The marker presented first (leftmost) is the one whose exclusion would have the strongest negative effect on the total collocation value of the set; put differently, it is the marker with the strongest contribution to the set. For the example in Table 7, the extracted set, French (NT Darby) 'know' is {-conn- | #sav- | #sach- | #sais# | #sait# | #sût#}.

A	B	C	D	E	F	**Columns**
1	1421.5	m	4	conn	244	A: Rank,
2	1405.7	m	5	#conn	229	B: Collocation value (Dunning's log-likelihood),
3	975.3	m	6	#conna	172	C: Candidate type (w = word-form, b = bigram,
4	974.2	m	5	conna	179	m = morph),
5	815.5	m	4	#sav	140	D: Number of characters of the candidate (word
6	815.5	m	3	sav	140	boundary is counted as a character; if the
7	620.0	m	4	onna	183	collocation value is the same, longer candidates
8	610.4	m	5	#sach	83	are ranked higher)
9	607.5	m	3	nna	183	E: Candidate (# stands for word boundary),
10	596.2	m	4	sach	83	F: Number of occurrences within search distribution
…						
47	368.0	w	6	#sais#	48	
…						
11717 candidates in total						

Table 6: Candidates ordered according to collocation value (for French and 'know').

A	B	C	D	E	F	G	**Columns**
1	1421.5	m	conn	244	331	1421.5	A: Rank,
5	815.5	m	#sav	375	504	2369.5	B: Candidate collocation value (Dunning's log-
8	610.4	m	#sach	455	593	3148.7	likelihood),
40	368.2	w	sais	493	635	3573.6	C: Candidate type,
80	174.4	w	sait	509	651	3782.1	D: Candidate,
330	16.5	w	sût	513	655	3837.8	E: No of verses with selected markers within search distribution (entire marker-set),
							F: No of verses with selected markers in entire corpus (entire marker-set),
							G: Set collocation value (Dunning's log-likelihood)

Table 7: Candidate selection for the same example as in Table 6 (for French and 'know'), only selected candidates listed.

Note that the most important part of the procedure is (c) Selection. Table 7 shows that the candidate with rank 330 is still selected (but totally only six of 11717 candidates considered were selected). A candidate with ranking number 330 would never be considered on its own. The only reason it is considered is that it is an asset when added to the set – unlike most other candidates.

For understanding how the algorithm works, it is further important to distinguish between *mutually dependent* and *mutually independent candidates*.

Word-forms are a very special kind of candidates in that their distribution is always mutually independent. For instance, the word-forms *sais, sait, savons, savez* all have their own, independent, sets of text occurrences. However, *savons* and *savez* are not independent of the potential morphs *#s*, *#sa, #sav*, *sav*, *sa*, *av*, *s*, *a*, *v*, whose sets of distribution all contain the sets of *savons* and *savez*. Selection and reevaluation are powerful for deciding which independent candidates to include or not to include. However, selection and reevaluation cannot easily handle the comparison of mutually dependent candidates. Once the algorithm has chosen the morph *#sav*, there is no way *savons, savez* can make it to selection, since the value for the set with them added will always be the same (having selected *#sav* includes them already). Once the candidate set contains mutually dependent items (which could be avoided by just having word-forms as candidates), ranking order becomes very important. In fact, "switching off" morphs for the example presented in Tables 6 and 7 has the effect of yielding a higher total collocation value, 4540.3 rather than 3837.8 with morphs "switched on" as candidates, as shown in Table 8.

However, this comes at the cost of a much larger marker set with thirty markers instead of just six and with a much lower coverage: equivalents of 'know' found in 453 verses rather than in 513.

The reason the collocation value is higher is that the marker set is better fitted – probably overfitted – to the specific search distribution. By overfitting we mean here that while the set adequately describes how the very specific search distribution ('know' in the American Standard English translation) in a specific text, the New Testament, can be matched, it is not necessarily the most representative for a more general 'know' meaning and for French in general. Forms included into the set only occur in 28 verses outside of the set (as opposed to 143 with morphs included). In this particular example, switching off morphs has the effect of making the set of markers more accurate for this particular text. Allowing for morphs, the strings, *-conn-*, *sav-* and *sach-* are actually quite representative for 'know' in the French text. However,

selecting them comes at the cost of including such word-forms as *connaissance* ‘awareness’, *saveur* ‘flavor’ and *savoureux* ‘tasty’. Including morphs makes the procedure less tightly fitted to a particular set of contexts in a particular translation and the result is more easily manageable. Five more general markers are a better summary than thirty very specific marker strings.

A	B	C	D	E	F	G
1	453.9	w	savons	43	43	453.9
2	426.9	w	connu	92	94	933.34
3	368.2	w	sais	139	145	1310.77
4	358.18	w	savez	181	190	1694.57
6	316.65	w	connais	208	218	1980.86
... sachant, connaître, sait, connaît, connaissez, sachez, savez-vous, sachiez, connaissons, connaissent, savait, sache, connue, connaisse, vous connaîtrez, ne connaissant, savaient pas, connaissait, connaissais, savais ...						
110	16.5	w	sût	441	469	4369.45
176	9.42	b	connaîtront que	444	472	4411.82
189	9.42	b	fais savoir	447	475	4454.42
192	9.42	w	connaissiez	450	478	4497.25
225	9.42	w	saches	453	481	4540.33

Columns: A: Rank; B: Candidate collocation value; C: Candidate Type; D: Candidate, E: No of verses with selected markers within search distribution (entire marker-set); F: No of verses with selected markers in entire corpus (entire marker-set); G: Set collocation value

Table 8: Candidate selection for the same data as in Table 6 (for French and ‘know’). Morphs (m) excluded (not all markers listed, since there are as many as 30 markers).

Table 9 shows the result with morphs included for several different French translations of the New Testament. As can be seen, there is a large degree of overlap despite the differences in the translations. Note in particular that the two leftmost markers (the most badly needed ones) recur across all translations. The results differ, among other things, in whether *conn-* has an initial word boundary (excludes forms of *reconnaître* ‘recognize’) or lacks it (includes *reconnaître*).

All translations happen to be quite far away from modern spoken French, but the The New World translation comes closest to what is expectable, reflecting also the past participle *su* and the stem of the future tense *saur-*, which occur too rarely in Darby and other translations to be extracted. The greatest variation can be found at the right border (close to the threshold value) where the results differ as to whether

forms of *ignorer* ‘be ignorant, not know’, *comprendre* ‘understand’ and *se rappeler* ‘recall’ make it to the set.

Excluding morphs is no option if the procedure is to be applied to all languages. In some languages with high morphological complexity (and in orthographies such as Japanese), word-forms (character sequences between spaces) are too rare to be retrievable one-by-one. Put differently, whereas including morphs does not always yield the mathematically best collocation value, our experiments with various sets of contexts across many languages have shown that it most often yields very good results and with a more limited number of markers than with word-forms only.

Translation	**Extracted marker set**	**Set collocation value**
darby	[conn | #sav | #sach | #sais# | #sait# | #sût#]	3836.7
perret	[conn | #sav | #sach | #sais# | #sait# | ignore | #sût#]	3691.1
nouvellesegond	[#sav | #conn | #sach | #sais# | #sait# | #saur | connaître | #reconnu]	3659.7
newworld	[#sav | #conn | #sach | #sais# | #sait | #saur | #su#]	3606.9
kingjames	[conn | #sav | #sach | #sais# | #sait#]	3428.1
jerusalem2004	[conn | #sav | #sach | #sais# | #sait# | #comprenez]	3335.9
ostervald1867	[conn | #sav | #sach | #sais# | #sait | #sût#]	3283.0
segond21	[#sav | conn | #sach | #sais# | #sait# | #saur]	3229.5
courant1997	[#sav | conn | #sais# | #sach | #sait | #saur | #rappelez-vous que#]	2556.0
paroledevie	[#sav | conn | #sais# | #sait# | #saur]	2506.5
semeur	[#sav | conn | #sais# | #sach | #sait# | ignore]	2388.1
despeuples	[#sav | conn | #sais# | #sach | #sait# | #saur | #ignor]	2360.0

Table 9: ‘Know’ across different French translations.

Why, then, keep word-forms as candidates? Couldn’t we just treat space as any character and provide all character sequences with, say, a maximum of twenty characters in length as one candidate set of potential “text morphs”? The reason is that word-forms are special in that they are mutually independent candidates, even though sequences of very different length. It is a functional advantage for identifying markers if at least some of them belong to a set of mutually independent candidates. We believe that this is a functional reason for why word-form is an important unit of language structure, reflected in many orthographies despite notorious difficulties of segmenting text into words.

The relevance of the Reevaluation step becomes particularly apparent if less powerful collocation measures are used, as already mentioned in Section 1. Replacing Dunning's log-likelihood with Dice (threshold 0.002) for the example shown in Table 6 yields *que* 'that' as first candidate (with morphs switched off). The selection step then results in: { *que connu sais savez sachant savons connais connaître sait connaît savait connaissent savaient connue reconnurent* }. However, the reevaluation step then reveals that the set value improves considerably if *que* 'that' is excluded from the set. Dice values range between 0.0 minimum and 1.0 maximum, and removing *que* makes the value rise from 0.314 to 0.763. With Dunning's log-likelihood, which is a much more accurate collocation measure, the first candidate very rarely needs to be excluded in Reevaluation. Testing the algorithm with a range of different collocation measures has convinced us that Reevaluation is necessary. However, the better the collocation measure to start with, the less Reevaluation has to compensate for its shortcomings.

3.3. The algorithm and otherwise associated items

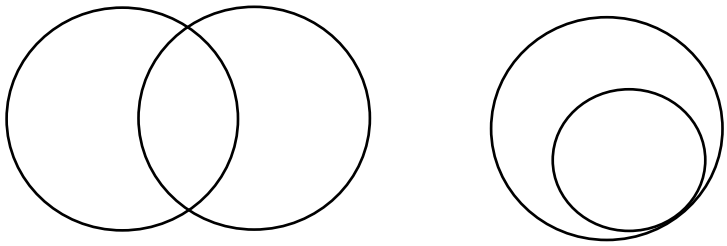
A crucial aim of the law and the algorithm instantiating it is to avoid mistaking otherwise associated items for markers (see Section 1). As certain kinds of otherwise associated items more easily slip through the net it is useful to classify them into rough types:

(i) *"Orthogonal associates"* have a large overlap, but mean something else, which makes them incompatible with certain contexts of the target meaning. For the meaning 'know', complementizers, negators and first person singular indexes are orthogonal associates since not all 'know' contexts have complement clauses, are negative or are first person. For negation, 'but' (contrast) is an orthogonal associate. In terms of sets, orthogonal associates are sets with considerable overlaps with the search distribution.

(ii) *"Partial associates"* go together with a subpart of the target meaning and take the form of subsets, but usually subsets that do not align well with individual markers. Partial associates can have special functions within the target meaning, such as negative polarity items, they can be emphatic reinforcers of the target meaning or they can be agreement markers of the target meaning. Reinforcers and agreement markers can be difficult to

distinguish from markers, and reinforcers can grammaticalize into markers (such as French *pas*, originally ‘step’, for negation), so here we have to expect a certain grey zone.

While both (i) and (ii) are mostly removed by the algorithm, difficulties arise if a partial associate aligns individually with a certain marker or with a subset of two or three individual markers and takes its place in the set instead. We will call this type of otherwise associated markers “shadows” and it will be illustrated in Section 4.



i) Orthogonal associates ii) Partial associates

Figure 3. Illustration of orthogonal and partial associates.

While reinforcers and shadows are expected to a certain extent as errors, orthogonal and partial associates can make it to the set of extracted markers if the true markers are not identified or if only some true markers are identified, which can be due to such factors as non-distinctive orthography, lack of segmental markers (the markers are not in the candidate set) or many rare suppletive or irregular markers. Many rare suppletive or irregular markers should not present a problem if the corpus is large, but in some cases the NT corpus is not large enough or not colloquial enough (as we already have seen in the case of the French past participle *su* ‘known’).

3.4. Sample

Typological investigations generally work with samples. It is practically impossible to investigate all approximately 7000 contemporary languages, notably because many of them remain insufficiently documented. If the population of interest is widened to include also extinct, future or possible languages (cf. Bakker 2010), total inclusiveness is not only practically but also principally impossible. Thus, a typological investigation requires some method of selecting a subset of the world’s known

languages for investigation. The selection may be done with different aims. Perhaps most common in typology is the aim of maximizing the linguistic variation found in the sample. This is known as a *variety sample*.

In this paper, we use the version of the Diversity Value method for variety sampling described in Sjöberg (2023). In the Diversity Value method, the focus lies on maximizing genealogical variation within the sample. This is done by applying an algorithm which turns the tree structure of classical language family classifications into a numerical value of complexity – the Diversity Value. The algorithm takes the number of branchings into account, but also the depth in the three at which the branchings occur; further-back branchings contribute more to the final Diversity Value. Languages in the sample are then chosen proportionally from the families based on Diversity Values (see Rijkhoff & Bakker 1998, Bakker 2010).

A problem with the Diversity Value method is that it offers no good way of choosing between families when the number of languages in the sample is smaller than the number of families in the given classification, which is often the case in typological investigations given that modern classifications contain around 250 families (e.g. Hammarström et al. 2023). Sjöberg (2023) therefore introduces a Diversity Value-based method which in addition to applying the Diversity Values algorithm also clusters families geographically. Families which do not have a sufficiently large Diversity Value to warrant inclusion in the sample on their own are grouped based on location (in addition, a logarithmic Diversity Value is used, to balance the role of very large families). The assumption is that just as genealogical variety correlates with typological variety, so does geographical variety. Thus, families which are geographically close are more likely to be alike than families far apart, allowing for the assumption that a language from one family in a group of geographically close families can represent the whole group. An additional reflection of the role of areality in the sampling method is the division of the world into five macro-areas, from which an equal number of languages are chosen. Unlike in some other approaches (e.g., Dryer 1989), languages are assigned to macro-areas based on their current location, but for simplicity's sake, families with only a very limited presence in one macro-area in terms of number of language (e.g. Indo-European in the Americas) are excluded in that area.

In Sjöberg (2023), the sampling procedure is applied to an as-complete-as-possible language catalogue, namely the Glottolog (Hammarström et al. 2023). This results in 19 empty sampling groups (of 95), i.e., groups which should be represented by a

language but for which there are no languages with a New Testament translation available. It would of course be possible to sample directly on the corpus catalogue – including only the languages for which there are translations available – but sampling based on the Glottolog allows us to see that there are 19 gaps (20%) in coverage as well as where these are.

As the Diversity Sampling method heavily relies on correct genealogical classification of languages, including languages with unclear affiliation is a challenge. Should, for instance, creoles be placed with their lexifiers, substrates, a family of their own or as isolates? Whatever choice made, it has considerable effects on the final sampling groups. The solution opted for in Sjöberg (2023) is to exclude creoles, creoloids and other languages with unclear affiliation from the core sampling and to add a small number of "wild card languages", which can also include historical languages, to the sample in the end. Here, Afrikaans, Middle English (historical language; enm; Indo-European, Germanic), Morisyen (mfe; French lexifier creole), Pennsylvania German (pdc; Indo-European; Germanic) and San Andres Creole English (icr; English lexifier creole) were added to the sample as an extension.

The entire sample consists of 83 languages (78 plus five wild-cards). See Appendix J for the list of languages.[14]

3.5. Getting started, with proper names

Let us first apply the procedure to proper names, since they can easily be evaluated manually and because proper names are expected to be translation-equivalent to a very high degree in parallel texts. We extract the markers for 'John' using the procedure described in 3.2 and the sample presented in 3.4. Examples are given in Table 10, for the full list see Appendix C.

Note that all forms are decapitalized; thus, the algorithm cannot see that proper names are usually upper case. Also note that the algorithm has no clue that we are looking for forms that are similar to *John*, *Johannes* or *Juan*. Further note that 'John' has strong associated items such as *baptizer,* none of which are wrongly extracted. Lemmatized Koine Greek (grc; Indo-European, Greek) *Ioannes* (Strong's number

[14] In one translation sampled, Doromu-Koki (kqc; Manubaran, New Guinea), 14 books of the NT are missing, but not the Gospels.

2491)[15] is used as meaning or search distribution (133 occurrences in 129 verses).[16] The log-likelihood threshold value used is 28.[17] Here as elsewhere, we have chosen thresholds with hindsight. We first try with a low threshold and test at which values wrong forms start occurring. Then we adjust the threshold so that it is just above that level and 28 is a low threshold.

In most cases, the result is entirely correct. If there is no inflection of proper names, the single word-form for 'John' is extracted (such as Igbo [ibo; Atlantic Congo, Igboid] *#jọn#*). If there is inflection, the longest shared letter sequence is extracted as a morph (such as Hungarian [hun; Uralic, Ugric] *#János*). In very few translations, more than one form is extracted, as in Toro So Dogon (dts; Dogon) { *#jan#* | *#jain#* } (the shared sequence *#ja* is no salient candidate).

Translation	Language	Set of markers	Recall	Recall (perc.)	Dedication	Set coll value
ibo	Igbo	[#jọn#]	127	98.45%	95.49%	1740.5
hun-revised	Hungarian	[#János]	127	98.45%	96.95%	1825.29
jpn-newworld	Japanese*	[ヨハネ]	127	98.45%	90.07%	1548.45
enm- wycliffe	Middle English	[#joon#]	125	96.9%	93.28%	1610.16
chr	Cherokee*	[#cani# | #canino# | #caniyeno#]	127	98.45%	95.49%	1740.5
tur-2009	Turkish	[ahy | #yuhanna]	120	93.02%	76.92%	1204.19
dts	Toro So Dogon	[#jan# | #jain#]	122	94.57%	71.35%	1163.12
kss	Southern Kisi*	[#chɔŋ#]	110	85.27%	63.58%	936.93
kmh-kalam	Kalam	[#jon#]	108	83.72%	62.79%	907.87
bvz	Bauzi*	[#yohanes]	116	89.92%	51.79%	894.89
eus-batua	Basque*	[#joan]	125	96.9%	27.29%	740.46

* Japanese (jpn, Japonic), Cherokee (chr, Iroquoian), Southern Kisi (kss; Atlantic-Congo, Mel), Bauzi (Geelvink Bay), Basque (Isolate, Europe)

Table 10: Markers for 'John' in the sample (selected languages).

It may come as a surprise that dedication (ratio of contexts that contain the expected marker which are within the search domain) is not close to 100% for many texts. This is because most translations use proper names more often (for co-reference instead of

[15] A system developed by James Strong in the 19th century (see Cysouw et al. 2007).

[16] English *John('s)* would be a less accurate choice since *John* sometimes also translates Ionas.

[17] In order to exclude Trinitario (trn; Arawakan, Southern Maipuran) *tvonicri'i*, probably 'baptizer', that would occur with value 27.4 (occurs in the 6 verses where *#Juan* is not used in the text).

personal pronouns) than the original Koine Greek text and translations that are close to the Greek original. Accuracy is almost perfect. In the Turkish (tur; Turkic) text, both *Yahya* and *Yuhanna* occur, extracted as { ahy | #yuhanna }, *-ahy-* because forms such as *vahyi* 'revelation' are wrongly included. (Not knowing yet that there also will be *#yuhanna*, the algorithm is a bit too greedy for the first form selected.) The very low dedication value for Basque is due to homonomy; *joan* is [go.INF], and could have been avoided without decapitalization.

In Figure 4, verses (x-axis) are ordered according to the number of sample languages with extracted markers (y-axis). Complete cross-linguistic identity would mean that in the 129 leftmost verses all 83 languages had extracted markers and then there would be zero languages in all other verses. Figure 4, where the result for the 400 top ranked verses (below there is almost only Basque *joan*) is given, shows that there is more diversity than might have been expected.

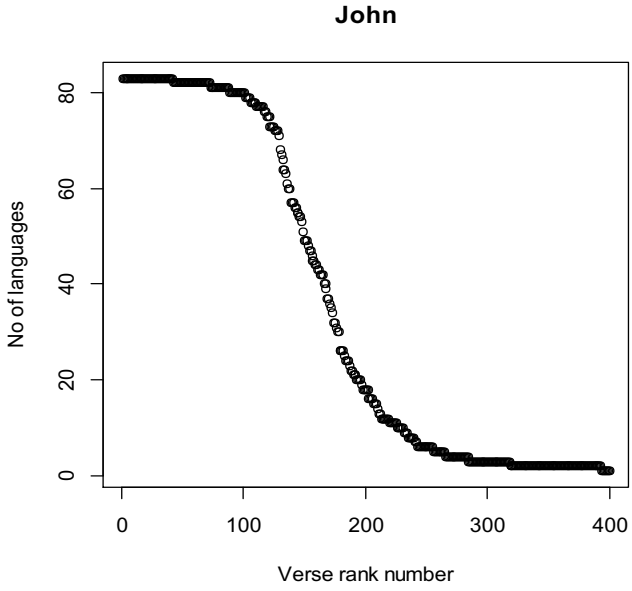


Figure 4: Occurrences of markers for 'John'.

Manual evaluation of verses without extracted forms in the 120 top ranked verses shows that most of the only 208 instances actually lack a form for 'John' (including many missing verses). Forms of 'John' missed by the extraction are all very rare forms, mostly hapax legomena: Middle English *Joones* (hapax) and *Joonys* (hapax) and Southern Kisi

Chɔ̄ŋ (3 times) – the algorithm cannot see that there is just a different diacritic here.[18]

The algorithm also works well with rarer proper names, such as *Herodias* with only six tokens in the search distribution in the American Standard English translation. With log-likelihood threshold value 21, the result is almost perfect (no wrong forms extracted, a form extracted in all languages of the sample). In six languages of the sample a bigram with an article or the like is best as Pennsylvania German *#di herodias#*, reflecting the fact that in these languages the name always or mainly occurs together with an article in the text.

4. Results and analysis

4.1. Introduction

Let us now apply the procedure applied to 'John' in 3.5 to negation (4.2), knowledge predication (4.3), first person singular subject (4.4) and propositional complementizers (4.5).

4.2. Negation

The sets of markers that our algorithm provides are a strong form of data reduction. In a result such as for Swedish (swe; Indo-European; Germanic) {#inte# | #ing | #aldrig# | #varken# | #förbjöd#}, there are unresolved abbreviations; *#ing* conflates *ingen* 'nobody' and *ingenting* 'nothing', the forms are not labelled; nothing in the set tells us that *#ing* stands for negative indefinite pronouns, *aldrig* for a temporal adverb ('never') and *varken* 'neither' for a negative connective. The constructions which the markers occur in are not accounted for (but note that the indefinite pronouns should only make it to the set if they are usually the single form of negation in the clauses where they occur, which is the case for Swedish). The most relevant marker is at the left edge (here the standard negator *inte*). Lexical negative forms, such as *förbjöd* [forbid.PST], can also occur, but if represented, will occur towards the right margin. However, lexical negative forms will not be systematically represented. It just happens to be the case that the past, but not the present, form of *förbjuda* occurs sufficiently often in the text considered in order to make it to the summary.

[18] Further examples are Dimasa (dis; Sino-Tibetan; Bodo-Garo) *jonthai* (one verse), Huitoto Murui (huu; Huitotan) *juandɨcue* (hapax), Purepecha (pua, Tarascan) *juanu* (one token with missing diacritic), Cherokee *canisgini* (hapax with additive clitic *=sgini*).

The set of markers is a descriptive summary similar to statistic measures such as mean value and standard deviation that summarize the properties of a set of numbers. We have verified all forms manually with the help of reference grammars, dictionaries and word lists, given in Appendix J. In the first column, the extracted markers are listed, the second column gives the manually added analysis.

swe-x-bible-2000 Swedish

inte	[NEG]
ing…	*ingen* ‘nobody’, *ingenting* ‘nothing’
aldrig	‘never’
varken	‘neither’ (in *varken…eller* ’neither…“or”’)
förbjöd	[forbid.PST]

The algorithm can be applied to texts in various writing systems and results may differ slightly due to writing system, such as for Kannada (kan; Dravidian, South Dravidian) when Latinized and in abugida – see Table 11.

Latin	lla#	abēḍ	āradu#	alār	rade	isad	ośśad	akūḍad	dilla	ārū#	#tiśiyad
Abugida	ಲ್ಲ#	ಬೇಡ	ಾರದು#	ಲಾರ	ರದೆ	ೊಳ್ಳದ	ೆಸದ	ದಿಲ್ಲ	ಕೂಡದ	ಾರೂ#	ಡದೆ#
Translit.	lla#	bēḍa	āradu#	lāra	rade	oḷḷada	isada	dilla	kūḍada	ārū#	ḍade#

Table 11. Extracted negation markers for Kannada in different orthographies.

Marker sets are only indirectly related to typological data points in typological databases such as WALS (Dryer & Haspelmath 2013). Thus, *bēḍa* happens to occur in the negative imperative (also called prohibitive), which is highly consistent with the classification “special imperative” in van der Auwera & Lejeune’s (2013) Prohibitive typology. The fact that all elements, especially the first one, are morphs rather than word-forms testifies to the value “Negative affix” in Dryer (2013). Our results do not reflect constructional features such as that standard negation in Kannada is asymmetric (Miestamo 2013). However, there is also partly more information than in WALS, notably concerning special markers for modal negation, such as *-bāradu* and *-kūḍadu* ‘must/should not’. The only form that should not have been extracted is *-ārū* (in *yār-ū* ’who-even’), which is a negative polarity item [NPI] that only occurs together with another negative form. Arguably wrongly extracted forms are given in red color in Appendix D.

kan-x-bible-latin Kannada

...lla	*-illa* [NEG, NEG.EX], *-alla* [NEG.COP]
...abēḍ...	*-bēḍa-* [PROH]
...āradu	*-ad(a)-* 'without', *bāradu* 'must/should not'
...alār...	*muchchalāraru* 'cannot close'
...rade...	mostly *bārade* 'must/should not'
...isad...	*-ad(a)* 'without'
...ośśad...	*-ad(a)-* 'without'
...akūḍad...	*kūḍadu* 'must/should not'
...dilla...	*-illa*: *iruvudilla* 'will not be'
...ārū	*yār-ū* 'who-even' [NPI] (very close to threshold 51.541)
tiśiyad...	*tiḷiyade* 'without knowing', *-adee* 'without'

To anticipate the general result, the algorithm performs very well for negation in terms of accuracy in that real negation markers are extracted for all languages of the sample and a clear majority of the extracted markers are indeed negation markers (black color in Appendix D). Coverage is respectable in that clearly in more than half of all relevant verses in all languages a negation marker was identified. No attempt was made to optimize recall for very rare markers. Rather, we use a relatively high threshold as each form extracted must be manually evaluated. Many negation markers mentioned in the descriptions consulted were not extracted and we did not evaluate whether this is because they are lacking in the NT or whether we missed them in the extraction. Notably, in cases of double exponence in negation, such as French *ne...pas* or Kaiwá (kgk; Tupian) *n(d)...i...*, usually only one of the syntagmatically co-occurring elements is extracted, which is expectable since all candidates in the algorithm (word-forms, morphs, and bigrams) are continuous strings. The issue might be addressed by allowing for discontinuous strings as candidates, which the present version of the algorithm does not.

Since information about in which verses of the New Testament negation is present irrespective of a particular language is not available, we begin with a negation search distribution defined by one marker in a language with a very broad general negation marker: Polish (pol; Indo-European, Slavic) *nie* (Biblia Gdańska) in 193 verses (237 tokens of *nie*) in the Gospel according to Mark. This is a parochially expressed meaning (negation in one language, Polish, with Polish idiosyncrasies). Using the algorithm, we extract 274 markers in the 83 languages of the sample (with log-

likelihood threshold value 31). Ranking all verses of the NT according to in how many languages an extracted marker occurs in descending order and cutting below 68 (1534 verses from the entire NT), we obtain an interlingua meaning distribution for negation (a sort of worldwide "interlingua negation") that can be expected to contain the most prototypical contexts for negation. The extracted 381 markers (log-likelihood threshold value 50) – 4.6 markers per language (all listed in Appendix D) – are manually evaluated with reference grammars and dictionaries.

Reapplying the interlingua negation distribution to Polish (although Polish is not in the sample), there are actually two Polish markers { #nie# | #ani# } – *ani* 'neither, nor'. Only five of 83 languages in the sample have merely a single extracted marker (6%). Unlike the name 'John' (4.4), negation is expressed by a set of several markers in a very clear majority of the languages of the sample.

Since the algorithm orders markers according to their importance, we can first consider the leftmost marker (the one first listed per language in Appendix D) and can conclude after manual evaluation that this is a negation marker in all languages of the sample. Let us now consider some languages where there are arguably issues with some of the extracted markers.

Since negation has many otherwise associated items (often called "negative polarity items" [NPIs]), we can expect that there is always *some* result, but manual evaluation is necessary for checking whether the extracted markers are negative polarity items. In particular, we can expect contrast markers ('but') and indefinite pronouns ('nothing, anything') to be wrongly represented in the result. Negative indefinite pronouns and negative adverbs (such as *never*) are acceptable in the result if the language does not have double negation such as Standard English, such as in the sample Middle English *neuer*... 'never' and Pennsylvania German *ken*... 'no' (see also Swedish above), but not in languages with double negation such as Afrikaans, where only *nie* is extracted (see Haspelmath 2013a).

Contrast markers are to be considered errors for the negation domain (orthogonal associate in 3.3). We can get them if too low a proportion of negation marker tokens were identified. There is only one contrast item ('but') among the extracted marker and this happens in the language with the most complex negation marking in the

sample: Yélî Dnye (Yele) [yle].[19] According to Levinson (2022: 495): "One of the most complex aspects of Yélî Dnye morphosyntax is negation [...] Essentially, the negative element fuses with the proclitic marking tense/aspect/mood/person/number in largely unpredictable ways, requiring rote learning." Yélî Dnye *ngmênê* 'but' comes up as third-ranked extracted marker. One way to eliminate it is to lower the log-likelihood threshold value to 20, then eleven other extracted markers push it out in reevaluation. Among these eleven strings, ten certainly occur in negation markers, the last lowest-ranked one is probably wrong (not attested in descriptions of Yélî Dnye). In total there are only few "Non-Described Forms" [NDFs] in the entire sample (forms that could not be verified with reference materials available to us), but a majority of them are clearly correct, judging from manual analysis of the forms in the texts.

Indefinite pronouns or similar elements which are negative polarity items (partial associates in 3.3) were wrongly extracted in Toro So Dogon, Kannada and Turkish. Two more languages are a matter of debate. Comaltepec Chinantec (cco; Otomanguean, Chinantecan) *jíi̱'˜ jaang*` [only one] 'nobody' appears to be a negative polarity item in the examples in the grammar, but occurs in some instances in the text as the single negation marker. Tlahuitoltepec Mixe (mxp; Mixe Zoque, Mixe) has *ka't* and the negative verbal prefix *ka-*; the latter is not extracted. But *ka-* usually co-occurs with an indefinite pronoun or adverb starting with <ni-> and without additional marking it is actually grammatical only if the negated constituent is the subject. It is thus not obvious whether the Tlahuitoltepec Mixe (mxp; Mixe Zoque, Mixe) verbal prefix *ka-* is to be considered a negation marker.

Several markers are ambiguous (homonymy or polysemy) and this is the source of a few errors if the non-negative item is more frequent than the negative one. Olo (ong; Nuclear Torricelli; Wapei-Palei) *pato* is a prohibitive marker, but *p-ato* is also [3PL-stay/be], which is why *turi* 'afraid' from *ise ma tur-ise pato* [2PL IRR afraid-2PL PROH] wrongly makes it above the threshold. Naro (nhr; Khoe-Kwadi, West-Kxoe) has a trigram *ta ga hãa* [NEG can/PARTICLE PST] with a rare negation marker *ta*, but *ta* is most often a pronominal index for first plural 'we'. This is why the bigram *ga hãa* (trigrams are no candidates in the present algorithm) makes it above the threshold. The string

[19] In a sense, Hungarian *nem#* also includes *hanem* 'but (contrast)' aside from standard negator #nem#. Our algorithm is too greedy in the beginning. When selecting that nem# is better than #nem# it does not know yet that it also will select #ne#, #se, incs# and #mégsem#. Actually, in the case of Hungarian negation, switching of morphs would yield a better total collocation value.

ga hãa is what could be called a shadow of the hidden (not extracted) marker *ta* [NEG] (see 3.3). A shadow is the "wrong" part of a very strong collocation pair in alignment with one or a few markers within the set of markers (see also 3.3). Further examples of shadow-errors are Cuiba (cui; Guahiboan) *dapo-* instead of *aibi/ajibi* (*dapon aibi, dapon ajibi* [DEM NEG.EX]) and Galibi Carib (car; Cariban, Guianan) *-iton* for the prohibitive forms *kytaiton*, *kysapyiton*, *kysupiton* with the very interesting Galibi Carib prohibitive markers *kyt-* and *kys-* that conflate inclusive (first and second person) affirmative with prohibitive (Courtz 2008: 88, 75).

Negation is in many ways an "easy" grammatical domain for our algorithm, because it is expressed in all languages. But the marking of it is not always salient in terms of invariant strings. However, in many languages, negation is synthetically marked in the middle of the verb, sometimes with a set of allomorphs. Thus, aside from the negative copula *değil-*, Turkish has the verb-internal standard negator *-mA-* where *A* stands for the vowel harmony variants {a, e}. Turkish *(-)ma(-)* and *(-)me(-)* also occur in many other non-negative uses, so they do not have high cue-validity for negation. The algorithm "solves" this by making a mosaic of less frequent elements { *... madı* | *maz* | *medi* | *miyor* | *meyece* | *mayın#* | *mıyor* | *mayaca* | *mama ... meyin#* | *mezs* | *emez ... rmeyen* | *mesin#* | *#korkma* | *masın#* } also containing following tense-aspect markers (such as *-iyor/ıyor* progressive) and participle or converb markers and occasionally preceding verb stems (*kork-* 'fear') or bits of preceding verb stems.

Since all candidates in our algorithm are segmental strings, only segmental markers can be found. What, then, if negation is expressed by reduplication, as in Hills Karbi (mjw; Sino-Tibetan, Karbic) (not in the sample) (consonant or consonant cluster from the verb stem + *e*)? Interestingly, even in Hills Karbi, more than half of the negation verses can be covered. The extracted set is { edet | #kali# | iri# | #chinine | #nangne }; *-edet-* is the /e/ from the reduplication plus the perfective suffix *-det*, *kali* non-reduplicative segmental copular negation. In addition, some frequently negated verb stems with their reduplication *chini~ne-* [know~NEG], *nang~ne-* [need~NEG] are extracted.

We can conclude that negation is generally well extracted with our algorithm. However, it is important to note that only markers are extracted, not negation constructions. Moreover, what is provided is a summary descriptive tool with strong data reduction.

4.3. Knowledge predication ('know')

We start with lemmatized American Standard English 'know' (598 tokens in 538 verses) as search distribution, a parochially expressed meaning, with log-likelihood threshold value 50, from which we derive an interlingua prototype with the sample languages, which we cut below 49 of 83 languages (59%) with extracted markers (536 verses in the NT). The interlingua version of 'know' is quite similar to the seed distribution, but lacks such idiosyncratic contexts as *know* in the Biblical sense (taboo expression for having sexual relationships). The result for English Lexham { *#kn* | *#recogniz* | *#ignorant* } is lexically not much broader, but also including 'recognize' and 'ignorant'. The Swedish result { *#vet#* | *#kän* | *#visste#* | *#veta#* | *#kunskap* | *förstå* } shows that the interlingua meaning verse set also includes part of the 'understand' domain (*förstå* 'understand') and that it contains what tends to be expressed by nominalizations (Swedish *kunskap* 'knowledge') in many European languages. English *knowledge* is also included in *#kn*, which – a bit greedily – summarizes *know(-)* and *knew* at the cost of wrongly extending also over *knee*, *kneel* and *knock*, which are, however, rare in the NT.

Using log-likelihood threshold value 50 has the consequence that some rare markers are missed. With threshold 40, more rare forms, such as Yélî Dnye *mya* 'recognize' and North Tanna (tnn; Austronesian, Oceanic) *iatun* [*ia-tun* DU-know] (an irregular dual form) would be included. A rather high threshold is chosen here for convenience in evaluation, since rare forms are often difficult to find in grammars and dictionaries. In total, 382 markers are extracted (4.6 per language on average).

The result is entirely correct in the sense that at least some markers for 'know' are extracted in all languages of the sample. It is not always verbs, as in Yélî Dnye { *ama#* } summarizing *lama* 'knowledge' and *l̠ama* [POSS.2.knowledge], illustrated in (6), where the marker is a noun with person marked in a possessive pronoun or prefix and occurring in a construction with an auxiliary proclitic and a positional verb (Levinson 2022: 334):

(6) Yélî Dnye (Yele; yle-x-bible, 43004025)

…A	***lama***	*ka*	*tóó,*	*yi*	*pini*	*dini*	
1SG.POSS	**knowledge**	CERTAIN.3PRS.CONT.IND	sitting	that	person	time	
ghi	*n:ii*	*ngê*	*wa*	*t:aa…*			
part	REL	ADV	3FUT.PUNCT	arrive			

'I know that Messiah is coming…' ("…when that person will arrive")

As can be seen in Appendix E, in many languages forms of several lexemes are extracted, which can make such distinctions as 'know person (*kennen/cognocer*)' vs. 'know fact (*wissen/saber*)' or lexically negative 'know' ('be ignorant') or 'know how/be able'.

In two languages, the first extracted marker is arguably wrong, because it is an otherwise associated item rather than the 'know' predicate, although of the reinforcing type (see 3.3). Ma'di (mhi; Central Sudanic, Moru-Madi) and Chol (ctu; Mayan, Cholan) happen to have very strong adverbial collocates of 'know' which are also highly dedicated to 'know'.

The Ugandan Ma'di adverb *òtē* '(know) properly; (see) well' (according to Blackings & Fabb 2003, a completion adverbial) requires a verb of perception or cognition (Blackings 2000: 83) and mainly occurs with *nì* 'know' in the NT. Actually, it occurs in most 'know' contexts, as illustrated in (7). That not many forms of *nì* <ni> 'know' make it to the extracted set of markers { ote |oniki | ini ta | anyini } is because of the not particularly distinct Ma'di orthography, which neither distinguishes tone nor /i/ vs. /ɨ/. There is a frequent pronoun *nī*, also written <ni>, and <ani> stands for both *á-nì* [1SG-know] and the much more frequent pronominal form *ānī* [3SG]. There is no way sufficiently many forms of *nì* <ni> 'know' can make it to the extraction to outrival <ote>. In a more distinctive orthography, the set of forms of Ma'di 'know' would together have a better collocation value than the adverb *ote*.

(7) Ma'di (mhi-x-bible, 43004025)

*A-**ni***	***ote***		*Mesia*	*ni,*	*ungwe-le*
1SG-**know**	**properly(PERC)**		Messiah	PRO	call-SUBORD
Kristo	*'i*	*ri,*	*k-e-mu*		*ra*
Christ	FOC	DEF	3DIR-VENT-go		AFF

'I know that Messiah is coming, the one called Christ'

In Chol, it is the adverbial *i sujm* <isujm> 'certainly, truly', which very often occurs in the 'know' domain, as illustrated in (8). Chol uses four verbs in the 'know' domain, *ujil*, *ña'ty* and *käñ* <cʌñ>, all meaning roughly 'know', and *ch'äm* <chʌm> 'take', which means 'understand' only when combined with *isujm*. The algorithm fails to extract *-chʌm-*, which is much less dedicated to 'know' than *isujm* 'certainly, truly'. Bigram candidates are no option since *-chʌm-* has too many different inflected forms. Only together with forms of *ch'äm* <chʌm> 'take' would the set of markers yield a

better collocation value if *isujm* ‘certainly, truly’ was omitted from it. In an ideal solution, *isujm* should be included only when combined with *ch’äm* ‘take’.

(8) Chol (ctu-x-bible-tili, 43004025)

*C-**ujil***	***isujm***	*mi*	*quejel*	*i*	*tyʌlel*
A1-**know**	**certainly/truly**	IPFV	start	A3	come.here

Mesías…
Messiah
‘I know that Messiah is coming…’

We can conclude that even though our algorithm cannot find the perfect solution for Ma’di and Chol, this does not invalidate the law of meaning discussed in this paper. It is just a practical difficulty, in Ma’di due to orthography, in Chol because an adverbial expression is an otherwise associated item in some, but part of a complex marker in other, contexts. But also note that Ma’di *òtē* and Chol *isujm* are otherwise associated items of the reinforcing kind (3.3), which come rather close to markers.

As mentioned in 2.3, it has been argued that there are languages, such as Kalam, that lack ‘know’. In Kalam, there is only a very general perception and cognition verb *niŋ-*. In our extraction, this stem is reflected in the two markers { niŋb | #niŋr } (*-b-* is perfect; Pawley & Bulmer 2011: 149). This does not necessarily confirm the view in Natural Semantic Metalanguage, that ‘know’ is a semantic prime (Wierzbicka 2018), but it shows that some sequences with *niŋ-* are sufficiently associated statistically with the ‘know’ domain that they can be said to express that meaning even though they also express many other meanings at the same time. Sjöberg (2023) finds that there are two languages in the sample that arguably lack ‘know’, both from New Guinea: Kalam and Fasu (isolate). Like Kalam, Fasu has a very general perception and cognition verb *hemakapuráka* ‘think, love, remember, know, understand’. Our algorithm finds [*#hemaka* | *himete* | *#asera*]; *himetēraka* is a lexically negative verb ‘ignorant of sth, not knowing, not understanding’ and *aserakā* ‘see, look, know (by seeing)’ is another very general perception and cognition verb. It is true that the Fasu and Kalam marker sets correspond to ‘know’ only to a limited extent and this is reflected in their low collocation values. The two languages have the lowest values in our sample (see Table 12).

Interestingly, as a general trend, languages from New Guinea and the Americas tend to have lower values than languages from Africa and Eurasia. This suggests that ‘know’ as modelled here by an only superfically interlingualized distribution (only

one step) is perhaps not yet a fully unbiased meaning that is equally adequate for all languages of the world. After all, we started modelling it with English 'know'. On the one hand, our algorithm finds expressions for 'know' in all languages of the sample, but, on the other hand, the match is not equally good for all languages of the sample.

Language	**Extracted markers**	**Verses**	**Coverage**	**Dedication**	**Coll. value**
Mandarin Chinese**	[i1dao* | ao3d | #ren4shi# | #qi3bu4 zhi1# | #ren4 de2# | #ren4 chu1#]	472	88.06%	82.95%	3682.07
Zarma**	[#bay#]	468	87.31%	83.13%	3651.21
Pennsylvania German	[#viss | #vays | #gvist# | #gekend# | #ich kenn# | #eisicht# | #unbekand]	450	83.96%	83.33%	3479.26
...					
Southern Nambikuára**	[a3la3kx | kǫ3nh | e3wxe]	209	38.99%	40.98%	643.28
Kalam	[niŋb | #niŋr]	323	60.26%	23.30%	596.07
Fasu	[#hemaka | himete | #asera]	442	82.46%	14.29%	490.98

* *i1dao* instead of *#zhi1dao* for 'know', because of greediness error (see Section 6.5)

** Mandarin Chinese (cmn; Sino-Tibetan, Sinitic), Zarma (dje; Songhay); Southern Nambikuára (nab; Nambiquaran)

Table 12: Languages with highest and lowest collocation values for 'know'.

4.4. First person singular ('I')

We pick English *I* (American Standard translation) as a starting point, hereby mimicking the bias toward European "non-pro-drop" languages in the syntactic literature dealing with grammatical relations. The Book of Acts is chosen, because in the Gospels it is mainly Jesus who is first person. The log-likelihood threshold value can be lowered from 54 to 36 after obtaining an interlingua distribution. The value 36 is chosen with hindsight; below that value, errors appear in several languages. While in the English seed distribution all examples in the set were English subjects (there are no examples with standard of comparison *than I*), with interlingua, English Lexham *I* drops to a coverage of 90.6%, but for most languages of the sample the coverage increases, which suggests that the start distribution was rather parochial for

English.[20] The interlingua distribution differs from English (and Koine Greek) notably in that it contains a few contexts such as (9), where English has NPs with possessive pronouns with body parts and emotional predicates, where the experiencer is first person singular.

(9) English (eng-x-bible-lexham, 44002026)
*For this reason **my heart** was glad and **my tongue** rejoiced greatly, furthermore also **my flesh** will live in hope* (44002026)

In freer translations, (9) tends to be rendered, for instance, as 'Then my heart is glad and **I** am happy. **I** will rest in hope'.

The selected languages listed in Table 13 show that the result is morphosyntactically very diverse across the languages of the sample.

The meaning first person subject (conflating transitive subject A and intransitive subject S) can, for instance, be primarily encoded by a subject pronoun (Swedish *jag*) or by ergative and absolutive pronouns as in Bauzi (*eho* ERG, *em* ABS). Chechen (che; Nakh-Daghestanian; Nakh) (*swo* <co> ABS, *as(a)* <ac(a)> ERG) is typologically similar to Bauzi, but has in addition experiencers with 'know' and related verbs in dative case (*suna* DAT).

In Japanese, the pronoun can bear topic (*watashi wa*) or nominative marking (*watashi ga*). In Turkish, the extracted marker is an index (verbal suffix *-m*). Tamasheq is dual in the sense that both the pronoun *năkk* and the index *-æɣ* are markers. In Warlpiri, the extracted marker is the subject second position clitic *=rna*, as Jelinek (1984) suggested for syntactic reasons. However, our result is entirely semantically, not syntactically, motivated. The subject second position clitic *=rna* just happens to be the most salient marker for first person subject in Warlpiri. In Culina, the dominant marker is the auxiliary form *o-na* [1SG-AUX]. Cashibo-Cacataibo is very interesting in that for first person singular subject, the marker extracted in the New Testament is *kana (<cana>)*, the first person second position clitic of the narrative paradigm (Zariquiey Biondi 2011: 484), as opposed to the form of the conversational paradigm

[20] An example of a verse with English *I* that is not included in the interlingua distribution is 44027034 *Therefore **I** urge* (Koine Greek: παρακαλω 1SG) *you to take some food…*, which in many translations is expressed without first person, as, for instance, in the Basque text *Jan, bada, mesedez…*, literally: "Eat, then, please…".

rina, which may be due to the predominantly narrative character of the New Testament. However, for second person singular and non-singular, the markers that would be extracted are personal pronouns – *min* [2SG.A] 'thou' *mits-* (*mitsun* 2DU.A, *mitsux* 2DU.S). Second person clitics in Cashibo-Cacataibo do not distinguish number, which does not make clitics salient for the meanings second person singular subject and second person non-singular subject. The example of Cashibo-Cacataibo shows that markers for different person-number values need not be in the same morphosyntactic slot. Rather, forms from different morphosyntactic positions extracted can reflect differences in patterns of syncretism (person-number coexpression).

Language	Extracted marker set	Verses covered	Coverage of set	Dedication of set	Colloc. value
Swedish (swe)	[#jag#]	126	84.56%	89.36%	864.86
Bauzi (bvz)	[#em# | #eho#]	139	93.29%	51.29%	439.74
Chechen (che)	[#ac# | #co# | #суна# | #aca#]	140	93.96%	80.00%	809.08
Japanese (jpn)	[わたしは | わたしが]	117	78.52%	92.86%	876.07
Turkish (tur)	[m#]	129	86.58%	43.58%	305.41
Tamasheq (taq)*	[#năkk# | eɣ# | săɣ# | yăɣ]	91	61.07%	75.21%	401.99
Warlpiri (wbp)	[rna# | rna-]	117	78.52%	65.00%	449.75
Culina (cul)*	[#ona | #ohuap]	126	84.56%	64.29%	495.58
Cashibo-Cacataibo (cbr)*	[#cana# | #'ëx]	128	85.91%	76.19%	649.83

*Tamasheq (taq; Afro-Asiatic, Berber); Culina (cul; Arawan, Madi-Madiha); Cashibo-Cacataibo (cbr; Pano-Tacanan, Panoan)

Table 13: Different kinds of encoding for first singular subject.

According to Van Valin (2005: 16), in head-marking languages, such as Tzotzil (tzo; Mayan, Tzeltalan), arguments are expressed by verbal affixes. If we look at different Mayan languages, which all are head-marking (Chol is the only Mayan language in the sample), the outcome is rather diverse. In Central Mam (mam; Mayan, Quichean-Mamean), it is indeed the ergative set affix *w-* that is extracted, and for Chol we get *ti-c-* [PFV-ERG.1-] and *c-* [ERG.1-], but Tzotzil (1997 translation) is mixed, with several verb forms (*j-na'* [1SG-know], *j-tic'* [1SG-put]) among the results, but also the pronouns *vu'un* [PRO.1SG], *vu'un=e* [PRO.1SG=FIN] and *cu'un* [POSS.1SG], and in Popti' (formerly called Jacaltec [jac; Mayan, Kanjobalan]), the extracted sequences *...ojan* (*-oj-an* [FUT-

FIN.1] and ...*han* reflect the first person sentence clitic =*an* (Day 1973: 57; Aissen 1992: 61), which can occur following each topic or sentence containing a first person singular or plural marker. Obviously, Popti' =*an* must be indirectly associated syntactically with first person subject, but it is still extracted as the most salient marker. Its form is more constant than the ergative first singular prefix with the allomorphs *(-)w-/(h)in-*.

At first sight, our method excludes true cases of Haspelmath's (2013b) "dual-nature view" where both pronouns and indexes are present throughout the entire domain. In our approach, one of them must be the marker, the other one an otherwise associated item of the meaning first singular subject. However, there is indirect evidence in favor of the dual nature view in that in some languages, the forms extracted can change completely from pronouns to indexes or from indexes to pronouns if the search distribution only slightly changes. Angor (agg; Senagi) is a case in point where the extraction listed in Appendix F picks a set of four different sequences reflecting indexes, whereas other attempts with only slightly different search distributions yield the first singular pronoun *ro* as single member of the extracted set. This suggests that first singular subject is different from negation in less clearly distinguishing markers from otherwise associated items.

The result for first singular subject is entirely correct in the sense of accuracy; all extracted strings or parts of it express first person singular and in forms that are functionally equivalent to subjects in English.[21] But coverage is often not close to 100%. In two languages, less than 50% of the verses are covered, in nineteen languages less than 75%. Pronouns are often better extracted than indexes, which is expected both because the seed distribution is pronominal and because indexes are less salient and often have various allomorphs. In Daga (dgz; Dagan), the second extracted form after *ne*, first singular pronoun, is the irregular suppletive stem *ang-* 'go (first person) as in *ang-en* [go.1-PST.1], *ang-in* [go.1-1SG]. As in Daga, extracted indexes can be verb-specific and extracted forms can go together with individual frequent verbs, as Yuracaré (yuz; Isolate, South America) *të-yle* [1SG.COOP-know] where the experiencer of 'know' is not expressed by the subject but by the cooperative object and *tütü-y(-)* [sit/be/stay-1SG.SBJ].

[21] One form in Comaltepec Chinantec is a shadow: ...*n'*... is a shortcut combining ...*n'̄n* and ...*n'́n*, first person being expressed by final =*n* [=1SG] after nasal.

Thai (tha; Tai-Kadai; Daic) is interesting in showing how text-specific our approach can be. Markers are determined not for the entire language, but for a particular text in that language. Thai has many personal pronouns, whose choice is dependent on such factors as "age, social status, gender, the relationship between the speakers, the formality of the situation and individual personality" (Smyth 2013: 42). Smyth lists as many as twelve forms that can stand for first person, only two of which figure in our extraction. The text we consider does not reflect the full range of factors that are relevant in Thai.

4.5. Complementizers

The extraction starts with Latvian (lav; Indo-European; Baltic) *ka* (for a description of Latvian complementizers, see Holvoet 2016) in the Gospel according to John with a log-likelihood threshold value 81.[22] In all attempts, knowledge predicates dominate to the extent that they must be accounted for in some way. From the extracted strings we selected those reflecting markers that do not mean 'know' for the interlingua distribution which results in markers from 37 sample languages. After assembling prominent verses again, which feature markers from at least 17 of the 37 languages, all verses where the lemma *know* occurs in the English Lexham translation have been removed, which yields a search distribution with 698 verses for the entire NT rather than 979 verses (28.7% with *know* removed). However, also other matrix predicates can be frequent, especially in languages with very general perception and cognition verbs such as Daga *anu-* 'hear', which in addition to removing 'know' verses necessitates a high threshold of 209 right above Daga *anu-* 'hear'. If such a procedure is followed, there is arguably full accuracy in the result even though there is a grey zone with verbally inflected or evidential quotative forms, which, however, are always in some way grammaticalized and not simply forms of a matrix verb 'say'.

The languages of the sample can roughly be classified into the following types:

(i) There is a complementizer and it is extracted, e.g.: Afrikaans *dat*, Basque *-ela*, Igbo *na* or Western Highland Purepecha *eska-*.

[22] We have experimented with several seed distributions with graphemically distinct declarative propositional complementizers such as German *dass*, Latvian *ka*, Estonian (ekk; Uralic, Finnic) *et* (also purpose clauses) and Hungarian *hogy* (English *that* does not work, because it is also a demonstrative) as well as sets of seeds from several languages.

(ii) There is no clear complementizer, but some forms, often non-finite, that frequently occur in complementation are extracted, e.g. Turkish *…duğu…* mainly represented by *ol-duğ-u-nu* [be-PTC.PST-POSS.3-ACC]

(iii) No form is extracted and there does not seem to be a complementizer, at least not with 'know'.

(iv) Some sort of quotative form is extracted: e.g., Olo *(ir)polo* 'say this, speak like', Hopi (hop; Uto-Aztecan, Hopi) *yaw* quotative.

(v) No complementizer is extracted, but there is one (seven languages): e.g., Comaltepec Chinantec *e* and Pilagá (plg; Guaicuruan) *da'* (see Appendix G for the full list).

A negative side effect of the high threshold is that no more than one marker per language is ever extracted. Secondary markers, as they occur, for instance, in Central Alaskan Yupik (esu; Eskimo-Aleut, Yupik), do not make it above the threshold.

Another interesting point is that a bigram is the best candidate in Meyah (mej; East Bird's Head, Meax), illustrated in (10). The word *oida* is an invariant complementizer derived from a speech verb (Gravelle 2004: 16), *rot* 'concerning' is a preposition.

(10) Meyah (43004025; see also Gravelle 2004: 225 for *rot oida* with 'know')

…Didif	*di-jginaga*	***rot***	***oida***	*Kristus …*	*em-en*
1SG	1SG-know	**concerning**	**COMPL**	Christ	IRR-come
si					
STATUS					

'I know that Christ is coming.'

The extracted markers are very diverse and vary highly in frequency. At two extreme poles, we can find the Hopi quotative particle *yaw* occurring in less than 10% of the search distribution and Matal (mfh; Afro-Asiatic, Chadic) *kà*, which is also a topic particle and "one of the most frequent free morphemes" (Verdizade 2018: 33), detected in more than 95% of the verses of the search distribution (but dedication is as low as 11%). Both markers only barely make it over the threshold.

4.6. Reconsidering which meanings considered are most relevant for the law

For demonstrating the relevance of the law formulated in this paper it is important that a substantial number of meanings are expressed by more than one marker. If

there is just one marker, we can dispense with the assumption that meanings are expressed by sets of markers. Moreover, the specific strength of our algorithm (finding markers that are not particularly salient by themselves) can only manifest itself if there are several markers. Finding one marker is actually nothing else than picking the candidate with the best collocation. Table 14 shows that the burden of proof is distributed rather unevenly across the meanings considered. It is the meanings with medium degree of difficulty that are most important for the law, represented here by negation, 'know' and first person singular subject.

	'John'	**Negation**	**'know'**	**1SG.SBJ**	**COMPL**
Average extracted marker per language	1.05	4.6	3.67	2.34	0.55
Ratio of languages with multiple markers extracted (errors not counted)	3.6%	94.0%	80.7%	66.3%	0%

Table 14: Comparing the meanings considered.

For proper names, we can get very far just with a good collocation measure. For propositional complementizers, it happens always to be just one that is found (a single salient one). Put differently, even if the algorithm works excellently even with proper names (and many nouns) and to a certain extent even for strongly grammatical meanings, it is verbs and universally expressed grammatical meanings that most strongly testify to its relevance, at least as far as the evidence so far surveyed suggests.

Some readers might object that we exaggerate number of markers by ignoring the notion of lexeme. However, in at least 69.9% of the languages, there are forms from more than one lexeme extracted for 'know', and a clear majority of the languages of the sample has forms of more than one grameme extracted for negation. Put differently, a good collocation measure would not do the job on its own even if all texts were lemmatized.

5. Discussion

5.1. Introduction

This section puts the results obtained into a larger context. 5.2 picks up some basic properties of the meaning–marker relationship that we have argued for throughout this paper and further considers what follows from these properties. Section 5.3

elaborates on one basic property listed in 5.2 – uniqueness of the meaning–marker relationship, which is perhaps most problematic in several respects. Section 5.3 also illustrates how the comparison of two similar meanings in our approach may relate to such traditional notions in semantics as (near-)synonyms and co-hyponyms. In Section 5.4 we turn back to semantics in general and discuss what approaches to meaning are compatible or not compatible with our approach. Section 5.5 turns back to the notion of coexpression. Section 5.6 addresses the issue of translation and, in particular, of using Bible translations as a data source. Finally, 5.7 discusses how the algorithm presented in this paper might be further improved.

5.2. Basic properties of the meaning–marker relationship and what follows from them

In this article we have rejected the canonical ideal of a one-to-one correspondence between meaning and marker and have argued that the meaning–marker relationship has the following properties:

(i) *one-to-many* (not one-to-one): a meaning is expressed by a set of markers
(ii) *approximate* (no full congruence): extensions of meaning and of markers are similar, not identical
(iii) *distributional* (rather than determined by convention): the meaning–marker relationship is reflected in discourse
(iv) *uniquely determinable* (despite a lack of one-to-one equation): there is just one optimal marker set per language corresponding to a meaning
(v) based on strength of *statistic association* (collocation): the optimal set of markers has the best collocation value
(vi) *general* (subject to the same law or mechanism for all meanings and for all markers): the same mechanism is at work for all meanings

Some consequences that follow from the properties listed are:

Markers in a set (i) expressing a meaning can be expected to be part of other marker sets expressing other meanings at the same time. *Several independent layers of information* (for instance, lexical and grammatical) can be stacked upon each other which allows for higher density of information in discourse than if relationships between meanings and markers would have to be strictly one-to-one.

Since the meaning–marker relationship is one-to-many (i), markers can be expected to group opportunistically to *coalitions* for optimizing the expression of

certain meanings. Lexemes are just one special case of coalition phenomena. Prominent meanings can be expected to be *attractors* for sets of markers.

Since the meaning–marker relationship is only approximate (ii), we can expect a high degree of *taxonomic flexibility*. Marker sets can be coalitions of hyponyms of the target meaning (e.g., 'know (fact)', 'know (person)', 'recognize' instead of 'know') without any need for postulating semantic atoms, or the marker set can express a hypernym of the target meaning (e.g., perception-cognition instead of 'know').

Since meaning–marker relationships must always be expected to be only approximate (ii), *coexpression* is the rule rather than the exception and there is no reason to treat certain kinds of coexpression in special ways.

Since linguistic categories highly differ in distribution (iii), identity requirements would entail an overarching categorial particularism. However, since markers only need be *similar* in extension to the meanings they express (ii), at least certain lexical and grammatical meanings can be said to be *expressed in all languages* despite large cross-linguistic diversity. Among those are negation, 'know' and first person singular.

The proposed law provides a universal mechanism (vi) to determine which set of markers in a particular language uniquely (iv) corresponds to a meaning, which makes it possible to *unambiguously establish meaning–marker relationships* even though there is no link of identity between meaning and marker, but only similarity (ii).

Since meanings are best described by way of distributional extensions (iii) and since distributional meanings cannot be expected to strictly conform to abstract semantic features, but are rather subject to family resemblance, it is hardly possible to define meanings extralinguistically. The best way to model cross-linguistically general meanings empirically is averaging over sets of markers in sets of as different languages as possible in parallel text corpora. This requires that going from meaning to marker (onomasiology) is preceded by a semasiological step (going from marker to meaning). This also implies that *semantic comparative concepts are not strictly extralinguistic*.

There are no predetermined slots where to look for markers, which entails a large amount of *morphosyntactic flexibility* in expression (also concerning parts of speech involved). Markers can be told apart from other items in discourse only due to statistical association (v).

5.3 Limits to uniqueness of results

In Section 4 we have always reported one result per feature and language, which suggests – as does the formulation of the law (2) – that the set of markers for a

meaning in a language is always strictly and uniquely determined. However, a result reported is just one measurement made under specific conditions (in a particular corpus, with a particular portion of the corpus, with a particular set of possible candidates, with an interlingua distribution derived from a seed distribution biased to one/a few particular language(s), chosen due to occurrence of markers per verse in a seed distribution in a particular proportion of a set of diverse languages, with a particular threshold for the collocation value chosen, using a particular collocation measure). Looking at the results from a single extraction as reported in Section 4 and the appendices does not make clear that some measurements are more stable than others. Put differently, for some features in some languages, small changes in choices made can completely alter the result.

This variability is thought-provoking in several directions.

First, as will be further discussed in 5.7, there is potential for improving the algorithm by optimizing the choices made. We are confident that the collocation measure chosen is well-motivated among those available, the parallel text corpus chosen has many shortcomings, but is the only one available suitable for our purposes, and considerable improvement can probably be made by further developing types of marker candidates.

Second, since many choices can be made where it is not clear whether any single solution is best, the question arises as to whether the meaning–marker relationship is really strictly unique. Sometimes, slightly different measurements will suggest that the correct solution alternates between two or several marker-sets that are nearly equally good equivalents of a meaning. In terms of subject markers, this corresponds to what Haspelmath (2013b) calls “dual-nature view” where both pronouns and indexes express subjects and a case in point discussed in 4.4 is Angor, where extracted marker sets sometimes are just indexes and sometimes just the personal pronoun for first person singular.

Third, the question arises as to whether possible choices might be deviations from comparing like with like and if yes, whether such deviations should be permitted or not. In extracting complementizers in 4.5, we subtracted the ‘know’ domain from the search distribution because ‘know’ is very strongly represented in complementation (at least in the NT). Further, we chose a very high threshold in order to avoid the extraction of any markers for perception or cognition predicates (see Appendix H for a summary of thresholds chosen). In dealing with knowledge predicates (4.3), no high threshold was chosen to exclude the extraction of perception/cognition hypernyms in Kalam and Fasu. In a certain way, thus, the result that knowledge predicates are

universally expressed in the sample whereas complementizers are lacking in many languages of the sample, is simply a consequence of different a priori choices made. Not removing the 'know' domain and using a low threshold for complementizers would have entailed the result that many languages express complementation by means of knowledge predicates and other cognition/perception predicates. This may seem counter-intuitive, but excluding these items is in a way a violation of our claim that we do not avoid certain particular types of coexpression when determining by which markers a meaning is expressed. It is beyond the limit of this paper to come to a neat conclusion about what is the correct thing to do and whether there is a single correct thing to do at all. However, it is important to note that both law and algorithm allow for considerable flexibility in outcome, especially via the level of threshold chosen. It is therefore important that choices made are reported together with the result. The specific choices made in each of the searches reported on here are summarized in Appendix H.

Fourth, the question arises as to what extent results are determined by initial seed distributions. We compared what happens when for negation Iu Mien *maiv* is chosen as a seed instead of Polish *nie* (with all other choices being the same as reported in 4.2).

Level	Type
1	Completely identical markers and the order of markers is exactly the same
2	Basically, all markers are the same, but potentially in different order or with slightly different morph borders (slightly different character sequences)
3	Same as 2, but only at least 2/3 of markers are basically the same
4	At least one marker is the related according to the criteria in 2
5	No similarity whatsoever

Table 15: Five (dis)similarity levels comparing the results of two extractions.

If we distinguish five rough levels of (dis)similarity as defined in Table 15 and presented in the notation 1:2:3:4:5 with increasing dissimilarity of type from left to right, the extractions based on interlingua distributions with Polish *nie* and Iu Mien *maiv* as seeds yields a (dis)similarity of 34:10:36:3:0 (or 41%:12%:43%:4%:0%). Put differently, a very high similarity of results in the 83 languages of the sample. In other words, the two different extensional sets for approaching negation are near-synonyms.

Compare to this the (dis)similarity profiles based on extraction of the two co-hyponyms German *kennen* and *wissen* (lemmatized for obtaining seed distributions, Luther-1912-version), no interlingua iteration added.

The cross-linguistic (dis)similarity summary here is 10:6:11:42:14, which means that while some of the languages have very different results, especially those 31 where 'know(person)' and 'know(fact)' are lexicalized differently (0:0:1:18:12),[23] there is a large number of sample languages, where the results are very similar, especially among those sample languages that colexify 'know(person)' and 'know(fact)' (note that among these are included languages which differentiate only a 'recognize' meaning, something which is fairly common): 10:6:10:24:2 and this even though there is almost no overlap in verses between the two different sets with German seeds.[24] The five levels are illustrated in Table 16:

Level	**Language**	**Seed *wissen*, 324 verses threshold = 20**	**Seed *kennen*, 62 verses threshold = 20**	**Dislexification 'kennen'/'wissen'**
1	North Tanna	[ɨtun | əruru#]	[ɨtun | əruru#]	No
2	Kalam	[#niŋb]	[#niŋbi]	No
3	Doromu-Koki	[#diba# | #toto#]	[#toto# | #diba# | #mama#]	No
4	Swedish	[#vet# | #visste# | #veta# | #känner dina#]	[#kän]	Yes
5	Mandarin Chinese	[#zhi1dao# | #xiao3de2# | #qi3bu4 zhi1#]	[#ren4]	Yes

Table 16: The five (dis)similarity levels of results illustrated.

The examples show how qualitative paradigmatic semantic relations such as near-synonyms and near-co-hyponyms with excessive cross-linguistic colexification relate to our quantitative approach.

[23] This includes two languages (Modern Standard Arabic [arb] and Middle English [enm]) where the distinction is somewhat different in being characterizable as a distinction between propositional knowledge and everything else rather than in most languages as a distinction between 'know (person)' and everything else.

[24] Excluding languages which dislexify 'recognize' as well as Southern Nambikuára (nab) which can be analysed either way yields 10:6:9:20:0.

5.4. What kind of meanings are we dealing with?

We have titled this paper “A law of meaning” without taking up reference, oppositions, concepts or definitions, which for many linguists are essential semantic units. So what kind of meaning are we dealing with?

The requirement that follows from our proposal is that any useful model of meaning must center around sets of discourse occurrences. This is the only requirement we have. Beyond this, meaning can manifest itself by way of rather different “senses” (other extensional and intensional models of a meaning), as sketched in Figure 5.

MARKERS <---->	RANGE OF MEANING <---->	OTHER MODELS OF MEANING
Set of markers	Set of discourse occurrences	(a) Set of similar exemplar uses (b) Set of referents in real world or modelled in possible worlds (c) Set of definitions such as paraphrases in an explanative dictionary (d) Set of discourse exemplars with graded membership (prototype and periphery) (e) One or several salient points in conceptual space (f) Set of oppositions to other meanings (g) Set of elements of various constructions (hereby granting membership to a set of constructions) (h) One or several profiles in image schemas (i) etc.

Figure 5: Towards a model of meaning.

The law presented is compatible with many different models of meaning; however, it does not require any specific item in the list of “other models of meaning”. It is also compatible with incomplete, diffuse, realizations of senses. For instance, there is no reason why (c) a set of definitions must be exactly congruent with a range meaning. Sets of definitions can make rough mosaics with “stones” approximately patching the extension of a set of meaning in the same way as we have shown that sets of markers in particular languages approximately cover them. Several authors have emphasized

social components of reference. According to Dewitt & Sterenly (1987: 49, mentioning Strawson 1959), reference is often borrowed. Speakers can "know" what they are talking about to different extents by way of referential chains.

However, what we have ruled out strictly is that meanings reflected in sets of markers are abstract concepts without any anchoring in language use. It is the anchoring in language use that is absolutely indispensable for any sort of meaning.

The concrete algorithm we use is dependent on attested occurrences. However, the law also applies to possible or probable occurrences (past, present and future). As formal semantics operates with reference in possible worlds, the law discussed here might be extended to possible discourse occurrences (to the extent this can be modelled, it is not implemented in this paper).

Anchoring in use does not necessarily entail a situational approach to meaning (Bloomfield 1933: 139; see Riemer 2010: 36). A parallel between sets of meanings and sets of situations only arises if markers are at the same time entire utterances (as may be the case with primary interjections and monomorphemic forms of greetings). As markers usually only are parts of utterances, individual markers mostly determine entire utterances to very little extent.

Finally, it is important to emphasize that the law described here is just one among different mechanisms at work in meaning. For instance, it does not say anything about how the meaning of markers relates to the meaning of combinations of markers. However, what we claim is that it is possible to address meanings of individual markers disregarding how they relate to meanings of combinations of markers or to meanings of their parts (which aligns well with construction grammar).

5.5. Coexpression and differentiation

The explicit study of coexpression requires the consideration of at least two meanings at a time, but the law suggested here and the algorithm implementing it only targets one meaning, ignoring all other meanings. Despite not directly addressing the problem of coexpression, we claim that our algorithm copes rather well with it. Coexpression in the case studies considered only rarely prevents the algorithm from establishing meaning–marker relationships. What we find is that shared expression has gradual effects. If Basque *joan* means both 'John' and 'to go', homonymy lowers the marker's dedication to 'John' (dedication is entirely gradual in our approach) and hereby the collocation value of the marker set, but *joan(-)* is still the optimal marker

for ‘John’ in Basque. In the same manner, it does not matter much for our law of meaning that the Kalam and Fasu words expressing ‘know’ also express other kinds of cognition and perception, but the values, when compared to other languages, show that the collocation is weaker. The algorithm is more strongly affected if the search meaning is rarely expressed and the other shared meaning is much more frequent. As we have seen, this may trigger what we call shadows; for instance, that the algorithm suggests to us that Olo *turi* ‘afraid’ is one of the markers for negation, because the rather rare prohibitive marker *pato* that *turi* ‘afraid’ goes together with is homonymous with the frequent form *pato* ‘they stay/are’. Our findings show that shared expression is no major obstacle for establishing meaning–marker relationships, which suggests that natural languages – as they indeed do – can work very well with considerable and widespread coexpression on all levels of lexicon and grammar. Earlier literature indicates that coexpression is limited rather by the conversationalists’ need of avoiding misunderstandings in communication, which, more specifically, constrains certain particular kinds of coexpression pairings (Gilliéron & Roques 1912; Gilliéron 1921; Xu et al. 2020).

As natural languages have a high tolerance for coexpression, they also have a high tolerance for polymorphy. However, our law suggests that polymorphy is constrained by *Mańczak’s Law of Differentiation* (see Section 1, Note 1), according to which irregular forms are never rare. Our algorithm cannot retrieve very rare markers. The findings in the case studies suggest that this very strong constraint does not prevent the algorithm from working well in *most* cases. However, we cannot find all markers for all meanings, at least not in the New Testament. As discussed in 4.2, the irregular French perfect participle *su* ‘known’ is too rare to be found in most French translations of the New Testament. This shortcoming might be simply due to the facts that the New Testament is too short a text for some markers and that the New Testament is a very specific (non-colloquial) text. However, the example very clearly illustrates how strongly the law suggested here is entirely dependent on discourse. We argue that the relationship between meaning and markers can only be established in language use. Language use is extremely variable, which entails that our law of meaning can be as important for the study of intra-language variation as it is for the study of cross-linguistic diversity. It just happens to be the case that this study has focused on linguistic typology and we have not discussed which kind of language use and how large an amount of text is required. These are all empirical questions that may be addressed by future research. However, we have shown that such a special and limited

text as the New Testament, and in many cases even just a smaller portion of the New Testament, is sufficient for demonstrating the general validity of the mechanism that we suggest. While the algorithm works rather well for most sample languages in all case studies, we have encountered some challenges for Mańczak's Law, notably negation in Yélî Dnye (5.2). However, whether such shortcomings are just a matter of limitations in corpus length or a more fundamental problem, we claim that our method has the potential of identifying the most problematic languages in a sample surveyed. Our results show that if you want to look at a language where the expression of negation is really complex, you should not fail to have a glance at Yélî Dnye, and if you are interested in whether 'know' is universal, you should have a look at such languages as Kalam and Fasu.

A somewhat surprising finding is that the algorithm would be able to cope with a much higher amount of suppletion in frequent forms than is actually attested in natural languages. When we designed the algorithm, we were surprised that it works perfectly well without any requirement of any sort of formal similarity between the different markers of the set. This means the law cannot explain why different markers used for the same meaning have a strong propensity to be formally similar and why analogic levelling is such a common diachronic process. Put differently, our findings suggest that the conversationalists' predilection for a high degree of transparency in the marker–meaning relationship cannot be explained by the law of meaning suggested in this paper. There must be other mechanisms that drive analogic levelling in natural languages.

5.6. Limitations of applicability and impact of translation effects

It may be argued that the mechanism described here is too limited in its application to be called a law. The availability of large chunks of text entails a written language bias, as spoken and signed language is not time-stable, but this is a shortcoming shared with other findings in quantitative linguistics and with corpus linguistics in general. Many linguistic generalizations can most easily be made in corpora. The application of the law is so far limited to translated texts, simply because we do not know how to appropriately define meanings fully explicitly in extensional terms if meanings are not modelled by way of other languages, if we want to avoid, or at least limit, bias towards particular languages. But that is a practical problem rather than a theoretical one. Finally, the choice of translations of the New Testament is motivated by our large-scale cross-linguistically comparative interest. Of course, the mechanism

could also be illustrated on a small set of European or Eurasian languages, but we wanted to show here that it also works well in languages that are maximally different from each other genealogically, areally and typologically.

Much work in typology is based on the abstract idea of translation equivalence. What we are dealing with here instead is real, actual, translations, ranging over a considerable spectrum of different translation strategies. Some Bible translations, especially older ones, are very literal. However, many Bible translations made after the Second World War have what de Vries (2007) calls a "missionary skopos" and are of the explicative type, which entails that they are much longer than the original. This can be seen, for instance, by the unexpected high occurrence of person name tokens (see 3.5 and Appendix C) in many translations to languages of the New World and the Pacific hemisphere. However, since we do not pursue an abstract ideal of one-to-one correspondence in translation equivalence, but use an optimality-based approach, it does not matter much for our application that different translations differ in extent of freedom of translation and in degree of explicativity. What can be affected are coverage and dedication values, which tend to be higher in literal translations.

What is most important, however, is that the meanings considered are amply represented in the corpus, which is one of the reasons why extraction with basic level concepts works better than with subordinate level concepts. All four domains considered in Section 4 are widely attested throughout the New Testament.

Finally, as we have seen in some concrete examples, orthography can be an issue, if it is not sufficiently distinctive. It does not matter much if orthography deviates from phonological representation, as long as the writing system remains distinctive. In 4.3 we have seen an example of how underspecified representation in Ma'di triggers a wrong extraction for the 'know' domain. However, also note that in some cases, writing systems and orthography can be more distinctive than phonology, for instance, in Mandarin or in Italian (ita; Indo-European; Romance) *e* 'and' vs. *è* 'is'.

5.7. The relationship between law and algorithm and how the algorithm might be improved

As argued in 5.6, our algorithm is most powerful if sets with more than one marker are extracted (and the law formulated in this paper emphasizes the paramount relevance of multiple marker sets). If we now consider how the algorithm could be improved, there is certainly some potential for improvement in which markers are extracted first. We have seen that the first marker extracted sometimes is too "greedy", meaning that a segment is picked that is too short just because there are

some rare forms that wrongly make the shorter sequence appear a better match, such as when Turkish *ahy* is picked instead of *#yahya* for 'John' or Mandarin Chinese *i1dao* instead of *#zhi1dao* for 'know'. This could be addressed by disqualifying candidates consisting of one frequent form and one or two hapax legomena. The matter is not entirely trivial, so we did not address it here in this programmatic paper, but there are certainly ways to avoid greedy sequences in a future improved version. A possible solution is that within a pair of mutually dependent markers the collocation value of the shorter one must exceed the collocation value of the longer one by at least the threshold.

More importantly, we should think about including subtraction when compiling marker sets. So far, our procedure is only additive. We consider candidates for inclusion in the set. But if we start with a very inclusive marker, we could test whether subsets of occurrences of strings containing the marker as a substring significantly better correlate with the contrary of the search distribution. To give a simple example, if the algorithm suggests that we should start with *#kn* for 'know' in English, there must be some way to subtract *#knee#*, *#kneel#* and *#knock#* because these sets of contexts included in the set *#kn* are no good match for 'know'.

A most obvious field with large potential for improvement is the types of candidate sets tested by the algorithm. For instance, if we already have bigrams (and we have shown that bigrams are relevant in some cases), we could now easily add, for instance, trigrams and "circumgrams" (trigrams with the middle word-form omitted). However, in this programmatic paper, we did not want to overdo it. Also, each new candidate type must be tested carefully. Adding a candidate type can eventually do more harm than good as each new candidate type adds a further potential source of errors. So far, all three candidate types included are continuous. However, we know that some markers are discontinuous. For instance, our algorithm will never find French *ne…que* for the meaning 'only'. Finding non-continuous markers and tackling non-concatenative morphology is a challenge. However, we have shown that we can get very far with just a few very basic segmental marker-sets. Adding further candidate types will produce some improvement, but will hardly change the picture fundamentally.

Each text example comes with its context and we have to decide about how much context is included. Here we have used rather large word windows, the verses of the New Testament. This works excellently where the meaning to be found is usually reflected only once in a verse, as is often the case for proper names and lexical meanings, such as 'know'. For negation, first person singular and, most markedly, for

complementizers, the result could probably be improved if word windows could be reduced to the level of the clause. Smaller word windows would allow for more focused searching.

The approach we have pursued here is that we model meanings (search distributions) stepwise. The underlying idea is that we can start with a parochially expressed meaning and then by extracting markers from a sample of languages with the algorithm arrive at a generalized distribution that more properly reflects the meaning we are looking for in a cross-linguistically representative way. Here we have – for simplicity – used the same sample both for modelling the interlingua meaning and for the extraction to be evaluated. This is, of course, not ideal; there is a risk of overfitting. We have also seen that, although the simple approach applied yielded quite good results, the results were not equally good for all languages of the sample. Modelling knowledge predicates starting from English *know* yielded on average quantitatively better results for languages of Eurasia and Africa than for languages of the Pacific hemisphere (indigenous languages of the Americas, New Guinea and Australia). In a way, this is a shortcoming. However, this result also suggests that our approach has considerable potential for identifying areal-typological differences in language use.

6. Conclusions

This study at the crossroads between linguistic typology and quantitative linguistics has a very basic and simple core message. We have argued that the relationship between meaning and marker can be described by a general law: *a meaning is expressed by the set of non-randomly recurrent markers that together are the best collocation of that meaning*, which makes it accessible to empirical investigation in parallel text corpora in a principled way. Our approach entails that it is profitable to view meaning extensionally (extensionally in discourse, not in the non-linguistic world of referents). To pair with meaning, markers cluster to sets. For lexical meanings, such sets can be lexemes, but lexemes and grammemes are nothing else but special cases of opportunistic coalitions of markers. Our approach can also accommodate phenomena of shared expression, such as coexpression (see 5.3), reflected as only gradually weaker match in terms of collocation value. For instance, general cognition and perception verbs in some languages of New Guinea, such as Kalam, can be markers of ‘know’ as much as knowledge verbs in Standard Average European languages; such markers just have

lower collocation values, but what counts as a marker rather than an otherwise associated item is determined by optimality: candidates being part of the set with the best collocation value within a language are markers. Accordingly, there are no strong requests for markers to be particularly dedicated to their meanings if only a marker is part of the marker set that is the best collocation of that meaning.

We have shown how the law can be implemented in an algorithm that works well for a range of different meanings including at least proper names, general basic verbs such as 'know' and generally expressed grammatical categories (negation and person) in languages with different genealogical affiliations and from different parts of the world. While the algorithm is entirely quantitative, the endeavor also requires traditional typological work, since in non-trivial cases extractions of marker sets must be evaluated manually.

Acknowledgements

We would like to thank two anonymous reviewers, Francesca Di Garbo and the members of the editorial board, Östen Dahl and Robert Östling for very many useful comments. We are also highly grateful to Dmitry Nikolaev for having suggested Dunning's log-likelihood to us as a collocation measure (Appendix I) and to Amanda Kann for helping us improve the pseudo-code in Appendix B.

Abbreviations

= = clitic
~ = reduplication
1 = 1st person
2 = 2nd person
3 = 3rd person
A = set A conjugation
A = transitive subject
ADV = adverbializer
AFF = affirmative
COMPL = complementizer
CONT = continuous
COOP = cooperative object
COP = copula
DEF = definite
DEM = demonstrative
DIR = directive
DU = dual
ERG = ergative
EX = existential
FIN = particle in final position
FOC = focus
FUT = future
INCOMPL = incomplete
IND = indicative
IPFV = imperfective
IRR = irrealis
NEG = negation
NDF = non-described form
NPI = negative polarity item
PERC = perception
PFV = perfective
POSS = possession
PL = plural
PRO = pronominal
PROH = prohibitive
PRS = present
PUNCT = punctual
REL = relative
S = (intransitive) subject
SG = singular
SBJ = subject
SUBORD = subordinate
VENT = ventive

References

Aissen, Judith. 1992. Topic and focus in Mayan. *Language* 68(1). 43–80. https://doi.org/10.1353/lan.1992.0017

Bakker, Dik. 2010. Language sampling. In Jae Sung Song (ed.), *The Oxford handbook of linguistic typology*. Oxford: Oxford University Press.

Beekhuizen, Barend & Maya Blumenthal & Lee Jiang & Anna Pyrtchenkov & Jana Savevska. 2023. Truth be told: a corpus-based study of the cross-linguistic colexification of representational and (inter) subjective meanings. *Corpus Linguistics and Linguistic Theory* 20(2): 433-459. DOI: 10.1515/cllt-2021-0058

Blackings, Mairi John. 2000. *Ma'di-English and English-Ma'di dictionary*. Munich: Lincom.

Blackings, Mairi & Nigel Fabb. 2003. *A grammar of Ma'di.* Berlin: Mouton de Gruyter.

Bloomfield, Leonard. 1933. *Language.* New York: Holt.

Courtz, Henk. 2008. *A Carib grammar and dictionary*. Toronto: Magoria.

Croft, William. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Cysouw, Michael, Chris Biemann & Matthias Ongyerth. 2007. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *STUF Language Typology and Universals* 60(2). 158–171. https://doi.org/10.1524/stuf.2007.60.2.158

Dahl, Östen. 2007. From questionnaires to parallel corpora in typology. *STUF Language Typology and Universals* 60(2). 172–181. https://doi.org/10.1524/stuf.2007.60.2.172

Dahl, Östen. 2016. Thoughts on language-specific and crosslinguistic entities. *Linguistic Typology* 20(2): 427–437. https://doi.org/10.1515/lingty-2016-0016

Day, Christopher. 1973. *The Jacaltec language.* Bloomington: Indiana University.

Dewitt, Michael & Sterenly, Kim. 1987. *Language & reality. An introduction to the philosophy of language.* Cambridge, MA: MIT Press

Diessel, Holger & Michael Tomasello. 2008. The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics* 12(2). 97–142. https://doi.org/10.1515/cogl.12.2.97

Dixon, Robert M. W. 2006. Complement clauses and complementation strategies in typological perspective. In Robert M. W. Dixon, & Alexandra Aikhenvald (eds.), *Complementation: A cross-linguistic* typology, 1-48. Oxford: Oxford University Press.

Dowty, David R. 1979. *Word meaning and Montague Grammar.* Dordrecht: Reidel.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292. https://doi.org/10.1075/sl.13.2.03dry

Dryer, Matthew S. 2013. Negative Morphemes. In: Dryer, Matthew S. & Haspelmath, Martin (eds.), WALS Online (v2020.3). Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/112)

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2013. The World Atlas of Language Structures (WALS) Online (v2020.3). https://doi.org/10.5281/zenodo.7385533 (Available online at https://wals.info)

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19. 61-74.

Evans, Nicholas & David Wilkins. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language* 76(3). 546–592. DOI:10.2307/417135

Firth, John Rupert. 1957. A synopsis of linguistic theory 1933-1955. *Studies in linguistic analysis*, 1-52. Oxford: Philological Society.

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Vanhove, Martine (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations,* 163–216. Amsterdam: Benjamins.

Georgakopoulos, Thanasis & Stephane Polis. 2018. The semantic map model: state of the art and future avenues for linguistic research. *Language Linguistics Compass* 12(2).

Gilliéron, Jules. 1921. *Pathologie et thérapeutique verbales*. Paris: Champion.

Gilliéron, Jules & Mario Roques. 1912. *Études de géographie linguistique: d'après l'Atlas linguistique de la France*. Paris: Champion.

Goddard, Cliff 2008. Natural Semantic Metalanguage: The state of the art. In Cliff Goddard (ed.), *Cross-linguistic semantics*, 1-34. Amsterdam: Benjamins.

Goddard, Cliff. 2012. Semantic primes, semantic molecules, semantic templates: Key concepts in the NSM approach to lexical typology. *Linguistics* 50(3). 711–743 https://doi.org/10.1515/ling-2012-0022

Goldberg, Adele & Suttle, Laura. 2010. Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(4). 468-477. DOI: 10.1002/wcs.22

Gravelle, Gilles. 2004. *Meyah: an east Bird's Head language of Papua, Indonesia.* Amsterdam: Vrije Universiteit Amsterdam. (Doctoral Dissertation).

Hartmann, Iren & Martin Haspelmath & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38(3). 463–484. doi 10.1075/sl.38.3.02har

Hammarström, Harald & Robert Forkel & Martin Haspelmath & Sebastian Bank. 2023. Glottolog 4.8. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.8131084. (Available online at http://glottolog.org, Accessed on 2023-12-05.)

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. Language, 86(3), 663-687.

Haspelmath, Martin 2013a. Indefinite Pronouns. In: Dryer, Matthew S. Haspelmath, Martin (eds.), WALS Online (v2020.3) Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/46)

Haspelmath, Martin. 2013b. Argument indexing: A conceptual framework for the syntactic status of bound person forms. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 197–226. Berlin: Mouton de Gruyter. DOI:10.1515/9783110331127.197

Haspelmath, Martin. 2023. Coexpression and synexpression patterns across languages: Comparative concepts and possible explanations. *Frontiers in Psychology* 14:1236853. DOI: 10.3389/fpsyg.2023.1236853

Haspelmath, Martin & Andrea D. Sims, 2010. *Understanding morphology*. 2nd edition. London: Routledge.

Holvoet, Axel. 2016. Semantic functions of complementizers in Baltic. In Kasper Boye & Peter Kehayov (eds.), *Complementizer semantics in European languages*, 225-263. Berlin: De Gruyter Mouton. DOI:10.1515/9783110416619-009

Horie, Kaoru. 1993. *A cross-linguistic study of perception and cognition verb complements: a cognitive perspective*, Diss., University of Southern California.

Jelinek, Eloise. 1984. Empty categories and non-configurational languages. *Natural Language and Linguistic Theory* 2. 39–76. https://doi.org/10.1007/BF00233713

Kehayov, Peter & Kasper Boye. 2016. Complementizer sematics – an introduction. In Kasper Boye & Peter Kehayov (eds.), *Complementizer semantics in European languages*, 1-11. Berlin: De Gruyter Mouton.

Levinson, Stephen C. 2022. *A Grammar of Yélî Dnye*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110733853

Liu, Y., Ye, H., Weissweiler, L., Wicke, P., Pei, R., Zangenfeind, R., & Schütze, H. (2023). A crosslingual investigation of conceptualization in 1335 languages. arXiv preprint arXiv:2305.08475.

Mańczak, Witold. 1966. La nature du supplétivisme. *Linguistics* 4(28). 82–89. https://doi.org/10.1515/ling.1966.4.28.82

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press. URL: http://nlp.stanford.edu/fsnlp/

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, 3158–3163. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf

McCarthy, John J. 2007. What is optimality theory?. *Language and Linguistics Compass* 1(4). 260-291. 10.1111/j.1749-818X.2007.00018.x.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197631

Miestamo, Matti. 2013. Symmetric and Asymmetric Standard Negation. In Matthew S. Dryer, & Martin Haspelmath, (eds.), WALS Online (v2020.3). Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/113)

Noonan, Michael. 2007. Complementation. In Timothy Shopen (ed.), *Language typology and syntactic description* 2: *Complex constructions*, 2nd edn. 52–150. Cambridge: Cambridge University Press

Pawley, Andrew. 1994. Kalam exponents of lexical and semantic primitives. In Cliff Goddard & Anna Wierzbicka (eds.), *Semantic and lexical universals*, 387-421 Amsterdam: Benjamins. https://doi.org/10.1075/slcs.25.19paw

Pawley, Andrew & Ralph Bulmer. 2011. *A dictionary of Kalam with ethnographic notes.* Canberra: Australian National University. http://doi.org/10.4225/72/56E977731EC84

Riemer, Nick. 2010. *Introducing semantics.* Cambridge: Cambridge University Press.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314. https://doi.org/10.1515/lity.1998.2.3.263

Rosch, Eleanor & Carolyn B. Mervis & Wayne D. Gray & David M. Johnson & Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, *8*(3). 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Saussure, Ferdinand de. 1967/8. *Cours de linguistique générale*. Édition critique par Rudolf Engler. 1-3. Wiesbaden: Harrassowitz.

Sjöberg, Anna. 2023. *Knowledge Predication – A Semantic Typology*. Ph.D. Stockholm University https://su.diva-portal.org/smash/get/diva2:1800727/FULLTEXT02.pdf

Smyth. 2013. Thai. *An essential grammar*. London: Routledge.

Strawson, Peter Frederick. 1959. *Individuals: An essay in descriptive metaphysics*. London: Methuen.

Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620904

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

van der Auwera, Johan & Ludo Lejeune. 2013. The Prohibitive. In Matthew S. Dryer, & Martin Haspelmath (eds.). WALS Online (v2020.3). Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/71).

Van Valin, Robert D., Jr. 2005. *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511610578

Verdizade, Allahverdi. 2018. Selected topics in the grammar and lexicon of Matal. Stockholm University MA thesis.

Vries, Lourens de. 2007. Some remarks on the use of Bible translations as parallel texts in linguistic research. *Language Typology and Universals*, *60*(2), 148-157. DOI: 10.1524/stuf.2007.60.2.148

Wälchli, Bernhard. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, 355-393. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110317558.355

Wälchli, Bernhard. 2024. We need world-wide corpus-based typology: A parallel corpus study of restrictives ('only'). *Travaux Neuchâtelois de Linguistique* 79: 69-157. https://doi.org/10.26034/ne.tranel.2024.4824

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. Linguistics, 50(3), 671-710. DOI 10.1515/ling-2012-0021

Wälchli, Bernhard & Sölling, Arnd. 2013. The encoding of motion events: Building typology bottom-up from text data in many languages. In J. Goschler & A. Stefanowitsch (eds.), Variation and Change in the Encoding of Motion Events, 77-113. Amsterdam: Benjamins. https://doi.org/10.1075/hcp.41.04w228l

Wible, David & Tsao, Nai-Lung. 2010. StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions. Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics https://aclanthology.org/W10-0804

Wierzbicka, Anna. 2018. I know: A human universal. In Stephen Stich & Masaharu Mizumoto & Eric McCready (eds.), *Epistemology for the rest of the world*, 215-250. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190865085.003.0010

Wittgenstein, Ludwig. 1958. *Philosophical Investigations*. Translated by Gertrude Elizabeth Margaret Anscombe. Oxford: Blackwell.

Xu, Yang & Khang Duong & Barbara C. Malt & Serena Jiang & Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition* 201. 104280. https://doi.org/10.1016/j.cognition.2020.104280

Zariquiey Biondi, Roberto. 2011. *A grammar of Kashibo-Kakataibo*. Melbourne: LaTrobe University. (Doctoral Dissertation).

Appendices

Appendices are available online at https://doi.org/10.5281/zenodo.10522345

Appendix A: Comparison to other approaches using parallel texts

In the present approach, we use entire Bible verses as information units. Cysouw et al. (2007) use smaller units based on simple cues in punctuation. Asgari & Schütze (2017: 116) use relative position within verses to reduce the size of information units. Large information units yield many possibilities for errors (all other words and character sequences in all verses in the search distribution), which puts the collocation component to the test. Many modern approaches use some kind of token-based word alignment (see, e.g., Beekhuizen et al. 2023: 438 and the literature discussed there) before a collocation measure is applied or instead of a collocation measure. It is unclear which approach is best and this may also depend on research aims. Token-based approaches are, for instance, preferable for determining word order relations (see Östling & Kurfalı 2023). However, Liu et al. (2023: §2) argue that using Bible verses as information units has the advantage of allowing for results beyond the word-level which is how "richer associations among concepts are obtained." For our purpose it is important to consider how well the collocation component performs when unaided by any sort of word alignment and, due to the theoretical relevance of our work, we cannot use any tools with black-box components such as neural networks.

While our approach is the only one to our knowledge that optimizes collocation values for sets of markers, there is, of course, other work with multiple extracted forms in a search. In token-based approaches, results can be different for each token. Liu et al. (2023), using Bible verses as information units, use iterated extraction, which means that once the best candidate is extracted, extraction continues with the smaller set of verses where the extracted marker(s) does/do not occur. Iteration is also used in Wälchli (2014) and Wälchli & Sölling (2013). Iteration entailing search distributions with highly varying size entail problems with determining collocation threshold values (Liu et al. 2023: B5), which is why Wälchli (2014) and Wälchli & Sölling (2013) use a suboptimal collocation measure, *t*-score, which it is less sensitive to search distribution size than others. Instead, Liu et al. (2023) use a coverage threshold (of 0.9), which seems to have a heavy impact on what kind of concepts the

approach is applicable to. The concepts they select are all nouns in English (Liu et al. 2023: A2) and nominal concepts tend to match much better than the verbal and grammatical concepts considered in this paper. Also consider in the results in Section 4 that coverage highly varies across concepts and languages and rarely reaches 90% with the concepts considered in our paper.

Most approaches have in common that they model meaning indirectly by way of choosing a form in another language, but differ in whether they account for the bias induced by the seed language(s) (Dahl & Wälchli 2016). Liu et al. (2023) model concepts by way of English forms, but then apply reverse search to find colexification patterns relative to English. Beekhuizen et al. (2023) start with English, but then use backtranslation to also include contexts that were not covered by English. Most comparable to our approach is Asgari & Schütze (2017: 113), who start with a seed (a "head pivot" "that is highly correlated with the linguistic feature of interest") which is then projected to a larger pivot set. However, our approach is less cherry picking. Rather than working with the languages where markers can most easily be found, we first define a diverse sample of languages to work with and then stick to that sample irrespective of how difficult or easy it is to work with it (3.4), which is more in the spirit of traditional typological methodology.

Additional references

Asgari, Ehsaneddin & Hinrich Schütze. 2017. Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages. In Martha Palmer, Rebecca Hwa & Sebastian Riedel (eds.). Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2, 113–124. Stroudsburg, PA, USA: Association for Computational Linguistics.

Dahl, Östen & Wälchli, Bernhard. 2016. Perfects and iamitives: two gram types in one grammatical space. *Letras de Hoje* 51(3). 325-348. https://doi.org/10.15448/1984-7726.2016.3.25454

Östling, Robert, & Kurfalı, Murathan. 2023. Language embeddings sometimes contain typological generalizations. *Computational Linguistics*, *49*(4), 1003-1051. DOI: 10.1162/coli_a_00491

CONTACT

bernhard@ling.su.se
anna.sjoberg@ling.su.se

Linguistic Typology at the Crossroads 4-2 (2024): 72-124 ISSN 2785-0943

On markedness in locative and existential predication: "Existential takeover", frequency and complexity in Siberian languages

CHRIS LASSE DÄBRITZ
UNIVERSITY OF HAMBURG
HEAD OFFICE OF THE GERMAN SCIENCE AND HUMANITIES COUNCIL

Submitted: 13/09/2023 Revised version: 3/04/2024
Accepted: 15/05/2024 Published: 23/01/2025

Abstract

The present paper investigates existential and locative clauses in fourteen Siberian languages. It is shown that all of them exhibit patterns of so-called "existential takeover", i.e. originally existential items occurring in locative predication. Starting from the observation that locative predication is frequently viewed as ontologically primary and functionally unmarked against existential predication, these existential takeover patterns are unexpected. Considering the text frequency and pragmatics of locative and existential predication, the paper argues that a markedness-based approach to these domains is unfeasible and leads to false predictions and generalisations. Following from this, it argues that a general typology of locative and existential predication must not contain any a priori restrictions regarding the observed linguistic realisations. Moreover, it proposes a two-layered design of such a typology which considers both the domains themselves as well as possible co-expression patterns.

Keywords: non-verbal predication; locative predication; existential predication; markedness; information structure; Siberian languages

1. Introduction

As widely known, the expression of locative and existential predications (*The book is on the table.* vs *There is a book on the table.*) is tightly interwoven in many languages

of the world. Often, both types share their morphosyntax entirely, only differing in word order, as, e.g. in Finnish (fin; Uralic, Finnic).[1]

(1) Finnish (Uralic, Finnic; personal knowledge)

a. *Kirja* *on* *pöydä-llä.*
book be.3SG table-ADE
'The book is on the table.' LOCATIVE

b. *Pöydä-llä* *on* *kirja.*
table-ADE be.3SG book
'There is a book on the table.' EXISTENTIAL

Although it is widely accepted that the linguistic expressions of locative and existential predications belong together, their relationship is far from settled. Most approaches – regardless of their theoretical framework – assume that locative and existential predications share their propositional content but may differ in their linguistic realisation which is mainly due to information-structural reasons (Lyons 1967: 390; Clark 1978: 87; Freeze 1992: 552; Hengeveld 1992: 94–100; Dryer 2007: 240–241; Creissels 2019: 38). In contrast, some authors, as, e.g. Milsark (1974) and McNally (2011), argue for a different propositional content. I take the former position in this paper, which I elaborate on more in Section 2.1. Section 2.2 wraps up shortly what has been done in linguistic typology regarding the expression of locative and existential predications.

Given this theoretical background, the paper investigates locative and existential predications in fourteen Siberian languages belonging to four language families (Uralic, Turkic, Tungusic, and Yeniseian) from a typological perspective. Section 3.1 describes the languages, their typological profile and the data used. Section 3.2 sketches the general affirmative patterns of the expression of locative and existential predication in the given languages, whereas Section 3.3 discusses their negative counterparts. One significant finding is that in all fourteen investigated languages, existential items are in some respect constitutive for locative predications. For example, Dolgan and Sakha (dlg and sah; Turkic, Northeastern) co-express existential

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

and locative predication using existential items in either type, like in the Dolgan examples in (2) (Däbritz 2022: 364–370).[2]

(2) Dolgan (Turkic, Northeastern; Däbritz 2022: 365)

a. *Bu karmaːŋ-ŋa-r möːčük* ***baːr.***
this pocket-POSS2SG-DAT/LOC ball **EX.3SG**
'There is a ball in your pocket.' EXISTENTIAL

b. *Onton ke biːr ogo-m Kiri͡es-ka* ***baːr.***
then well one child-POSS1SG Kresty-DAT/LOC **EX.3SG**
'Then one of my children is in Kresty.' LOCATIVE

Similarly, negative locative predications exhibit existential patterns in all investigated languages. As a case in point, Kamas (kms; Uralic, Samoyedic) uses the negative existential item *naga* ~ *nago-* in negative locative and existential clauses (3a, 4a) (Däbritz & Wagner-Nagy 2024: 10-12).[3] In turn, affirmative locative and existential clauses (3b, 4b) display the non-existential copula verb *i-* (ibid.).

(3) Kamas (Uralic, Samoyedic; INEL Kamas Corpus: PKZ_196X_AngryLady_flk.044, PKZ_196X_SU0203.PKZ.071)

a. *Da tăn gijen-də i* ***nago-bi-al.***
and 2SG where-INDEF and **NEG.EX-PST-2SG**
'But you haven't been anywhere.'

b. *Ši? d'ije-gən* ***i-bi-le?.***
2PL taiga-LOC **be-PST-2PL**
'You were in the taiga.'

[2] As pointed out by both anonymous reviewers, the classification of *baːr* as an existential item needs justification. This issue is targeted in detail in Section 3, dealing with Dolgan and Sakha *baːr* in Section 3.4.1. In a nutshell, the argument is as follows: from a synchronic perspective, the item *baːr* is used in generic existential clauses like *God (does not) exist(s)*, it carries the meaning 'presence; existence' when used nominally, and from a comparative perspective, the item has cognates all over the Turkic language family, in most languages being restricted to existential (and possessive) predications.

[3] Again, a justification for the classification of *naga* ~ *nago-* as "existential item" is needed, as correctly pointed out by the anonymous reviewers. In this particular case, the item again appears in generic existential clauses, and an aspectual derivation can yield a meaning 'disappear', both pointing to an existential meaning of the item. See Section 3.4.2 for details.

(4) Kamas (Uralic, Samoyedic; INEL Kamas Corpus: PKZ_1964_SU0207.PKZ.094, PKZ_196X_SunMoonAndRaven_flk.004)

a. *Maʔ-na-l* *sazən* ***naga.***
tent-LAT-POSS2SG paper **NEG.EX.3SG**
'There is no paper at home.'

b. *A* *bărа-gən* *ši* ***i-bi.***
and sack-LOC hole **be-PST.3SG**
'And there was a hole in the sack.'

In either case, it can be shown that the relevant items indeed have initial existential semantics, so they have been taken over from existential to locative predication. To account for the initial existential semantics of the relevant items, Section 3.4 discusses their synchronic behaviour, their diachronic sources and related issues.

Often, it is at least implicitly assumed that locative predication is ontologically primary against existential predication; see, e.g. Lyons (1967: 390) and Freeze (1992: 554–555). Creissels (2019: 41) even explicitly states locative predication (in his terminology: *plain-locational predication, PLP*) is unmarked, and existential predication (in his terminology: *inverse-locational predication, ILP*) is marked. Markedness relates here to functional-semantic, not yet to formal aspects, so more precisely, these accounts assume that existential predications are functionally marked against locative predications. Since the functionally unmarked item in a markedness opposition is expected to spread or being generalised instead of the functionally marked item when the formal opposition of the items is neutralised (Greenberg 2005[1966]: 28–29; Waugh & Lafford 2000: 275; Bybee 2011: 134–135), the "existential takeover" patterns shown above are not expected. This direction of generalisation is frequently explained by the unmarked item being neutral for the whole category expressed – e.g., a present tense item can often be used in semantically past-tense contexts, but not vice versa – so locative predications would be expected to possibly appear in existential contexts, but not vice versa.[4] Consequently, an explanation is needed why

[4] Thanks are due to an anonymous reviewer who pointed me to the fact that diachronic syntax indeed exhibits many instances of loss of markedness or markedness reversal. This observation is surely correct, as spelt out by, e.g., Janda (1996: 215–217). However, to the best of my knowledge, such cases mostly relate to formal markedness (e.g., loss of formal complexity in paradigm regularisation) or usage-based markedness (e.g., replacing the genitive case in German with *von*-PPs) instead of functional-semantic markedness. In any case, the argumentation will show that a markedness-based

some languages still use existential items to express locative predication. In Section 4, I discuss the relevant issues of markedness, formal complexity, salience and frequency. I argue that the observable higher text frequency and formal complexity of existential clauses are the prerequisite and outcome of their higher degree of salience, respectively. Following this explanation, I argue that the notion of markedness has no explanatory force when applied to locative and existential predications since it leads to incorrect expectations and faulty generalisations.

Section 5, finally, points to some immediately following typological implications, which mainly target the design and structure of a general typology of locative and existential predication. Section 6 ties loose ends together and gives an outlook on related questions, including unsolved issues calling for further research.

2. Locative and existential predication

2.1. Delimiting the domain

In this paper, I conceive locative and existential predications from a functional-semantic point of view as expressing the presence or absence of a figure (a.k.a. theme, pivot) in a ground (a.k.a. location, coda). For terminological clarity, I distinguish "locative/existential predication" for talking about semantics and pragmatics from "locative/existential clauses" for talking about linguistic structures and morphosyntax. In this context, it is worth noting that Martin Haspelmath (p.c.) pointed me to the problem of the predicability of existential clauses: following Croft (2022: 290–293, 304–305), the term *existential predication* is a misnomer since existential clauses are per default thetic and, thus, non-predicational clauses in his framework. As shown by Sasse (1987), among others, thetic sentences indeed do not include a concrete referent, about which something is predicated. Still, I assume that it may also be the temporal/local circumstances of a situation in general, which may be the reference point for a predication, called *contextual domain* by Francez (2007: 70–71). Take, for instance, the English existential clause *there is no more coffee*. World knowledge and assumingly also the (extra-)linguistic context suggest that it is not meant to say that there does not exist any coffee at all. Instead, the speaker intends the reading that there is no more coffee to drink in a given situation. Thus, the

approach to locative and existential predication is unfeasible since it leads to contradictory expectations.

existential clause refers to this situation, its contextual domain. From an information-structural point of view, such references have been labelled *abstract topics* (Junghanns 2002: 45; Däbritz 2021: 97–98) or *stage topics* (Erteschik-Shir 2019: 233–235), their linguistic realisations showing several peculiarities, e.g. pitch accent on the subject, verb fronting, among others (ibid.). Consequently, I assume that also thetic sentences may count as predications – though not being their classical representation – which is undoubtedly relevant for the following description of existential predications.

As shown by Hengeveld (1992: 96–98), Koch (2012: 538–541, 545) and Haspelmath (2022: 17–20), the prototypical instances of locative and existential predication are clauses like (5) and (6), respectively.

(5) *The book is on the table.* LOCATIVE

(6) *There is a book on the table.* EXISTENTIAL

Either type of predication expresses location and not the mere existence of a referent, which is why they are often subsumed under one umbrella term such as *locational construction* or alike (Hengeveld 1992, Creissels 2019, Haspelmath 2022, among others). Predications, which lack a specified location (7), represent a different, though often formally similar type of predication (Koch 2012: 538–541, 545; Creissels 2019: 44–45; Haspelmath 2022: 17–20). Following Koch (2012), I call them *generic existentials*.

(7) *There are many unhappy people.* GENERIC EXISTENTIAL

Whether or not sentences like (7) can be discussed and analysed together with sentences like (5) and (6), thus belonging to the same functional domain, cannot be discussed in detail in this paper. Therefore, I leave them out of the systematic discussion and limit the core analysis to locative and existential predications containing a concrete reference to a location. However, as suggested by an anonymous reviewer, I take them into consideration when proving the existential semantics and existential origin of an item under discussion. Due to the non-expressed ground element, generic existentials are less close to locative predication than locational existentials. Thus, it can be expected that semantically existential items appearing in locative and locational-existential clauses must also appear in generic existential clauses (e.g., Dolgan *baːr* as discussed above). In turn, non-existential items can be restricted to locative and locational-existential clauses, as opposed to generic

existential clauses (e.g., the bare English copula verb *be*: *The book is on the table* vs *On the table is a book* vs **Books on linguistics are*).

Following Hengeveld (1992: 94–100) and Creissels (2019: 37), among others, I assume that locative and existential predications have the same propositional content, and their difference lies in the perspectivisation of the relationship of figure and ground. The terms *figure* and *ground* go back to Talmy's (1983) seminal work on the linguistic structure of space. Whereas the figure is a movable referent whose site, orientation etc., are variable, the ground is the reference object for the site, orientation etc., of the figure (Talmy 1983: 232). As for *perspectivisation*, Borschev & Partee (2002) operationalise the term via presupposition, assuming that the perspectival centre of an utterance must be presupposed in a discourse. Thus, in locative predication, the perspectival centre of the utterance is a presupposed figure referent, whereas it is a presupposed ground element in existential predication. In terms of film language, locative predications thus provide a close-up view of the figure, whereas existential predications provide a total view of both the figure and ground.

Indeed, the cognitive perspectivisation of a predication and its linguistic expression are rather abstract and often hardly observable in linguistic structures. However, it is reflected in the information structure of an utterance, more precisely, in its focus-background structure. The focus-background structure of a clause expresses what is most important for the speaker in the given context and what the speaker wants to emphasise (Molnár 1991: 58; Junghanns 2002: 13). This approach is in line with Lambrecht's (1994: 207) assumption that "[…] focus is what makes an utterance into an assertion" since the speaker contributes important (e.g. new, unexpected, correcting) information to the communication to bring the latter forward. So, in the case of locative and existential predications, the speaker either emphasises that the figure is somewhere (locative) or that there is a figure somewhere (existential). In more technical terms, the figure element must not be included in the focus domain in locative clauses, but it is necessarily part of it in existential clauses.

According to Hengeveld (1992: 119–120), existential predications are presentative constructions since they (re-)introduce a referent – the figure element – into the discourse; in terms of functional grammar, existential predications are thus [+presentative], whereas locative predications are [-presentative]. From an information-structural point of view, existential predications thus correlate with sentence focus, which is why the figure is necessarily included in the focus domain of the clause (Lambrecht 1994: 179; Bentley et al. 2015: 47–48; Erteschik-Shir 2019: 233).

Often, it is assumed that this entails the figure being indefinite per default in existential predication, referring to Milsark's (1974) *definiteness restriction*. As convincingly shown by Borschev & Partee (2002: 116–117) and Creissels (2019: 48–49), the correlation of indefinite figures and existential predication holds as a tendency, but not as a condition, cf. the Russian (rus; Indo-European, Slavic) example (8).

(8) Russian (Indo-European, Slavic; Borschev & Partee 2002: 116, glossing adapted)
Context: I was looking for kefir in the shop.
Kefir-a *v* *magazin-e* *ne* *by-l-o.*
kefir-GEN in shop-LOC NEG be-PST-N
'There was no kefir in the shop.'

Here, we see the seemingly contradictory properties that (a) the figure *kefir* 'kefir' is aforementioned in the immediate left context of the clause, but (b) sentence focus – answering the heuristic question of *what happened (then)?* – still yielding an existential reading. In terms of information structure, existential predications thus do not exhibit a segmented focus-background structure, as typical for thetic sentences (cf. Sasse 1987), regardless of the semantic-pragmatic properties of the figure element included. In contrast, locative predications have both a segmented topic-comment and focus-background structure: the figure functions as the topic of the clause, but more importantly, it is presupposed, backgrounded and, thus, excluded from the focus domain (Däbritz 2021: 146–147). The ground, in contrast, is included in the focus domain, the latter being either predicate focus or argument focus in Lambrecht's (1994: 226–233) terms.

In a nutshell, the main distinction between locative and existential predications is their cognitive perspectivisation, which results in non-presentative predicate/argument focus structures in locative predication. In contrast, existential predications are characterised by their presentativity, linguistically expressed by sentence focus structures. In the former case, the figure element must not be part of the focus domain, but in the latter case, it is.

2.2. Typological approaches

Locative and existential predications have been dealt with from various perspectives, including typological approaches. Still, a general typology targeting one or even both of them is yet missing. Assumingly, this is no coincidence but can be explained by the complexity of the domain(s) on the one hand, but even more by the unsolved questions of what to include in the domain and whether we are dealing with one or two domains. In what follows, I try to wrap up existing typological approaches, showing their benefits and caveats, and point to several issues important for this paper.

The first systematic typological approach to locative and existential predication is provided by Clark (1978), explaining word order patterns in locative, existential and possessive predication in roughly 30 languages. Starting from the assumption that the "configuration" of locatives and existentials is shared, Clark (1978: 94–96) argues that definiteness, instantiated in word order permutations, differentiates locative from existential readings. Since Clark (1978: 89–90) assumes that the shared configuration includes locative features, the approach implies that location is ontologically primary to existence. In other words, referring to Kahn (1966) and Lyons (1967), it is argued that existence presupposes location (ibid.). The latter assumption is shared in many subsequent works like Freeze (1992), Hengeveld (1992), Koch (2012) and Creissels (2019), among others.

Whereas functional aspects of locative and existential predication (e.g. Hengeveld 1992) and syntactic accounts to word order permutations in them (e.g. Freeze 1992) took this as their starting point, the morphosyntactic expression of locative and existential predication lacked an in-depth analysis. Stassen (1997) undertook the task of developing a typology of intransitive predication, whereby his approach was deliberately limited to non-presentative intransitive predications (verbal, nominal, adjectival and locational) with a definite subject NP (Stassen 1997: 9–10). Consequently, existential predications are not covered, but the expression of locative predications got some insightful treatment. Stassen (1997: Ch. 2 & 3) singles out three strategies (verbal, nominal, locational) for expressing the above-mentioned types of intransitive predication. The verbal strategy uses bound person-number-gender markers attached to the predicate; the nominal strategy uses an overt or covert copula (which eventually agrees with the subject in person, number and gender); the locational strategy, finally, uses a locative verb agreeing with the subject NP (Stassen 1997: 34–35, 55, 91–95, 111). Apparently, the "default" case is that the verbal

strategy expresses verbal intransitive predication, et cetera, and the crucial point of interest for applying the developed typology is the notion of "strategy takeover". If a language uses a strategy, which is not prototypical for the relevant type of predicate on the synchronic level, the language is assumed to take over the strategy under discussion (Stassen 1997: 29–30). Evidently, this notion is central to the paper at hand since it analyses instances of "existential takeover" in locative predications, which means that an existential strategy is applied to a non-presentative locative predication.[5]

In the realm of locative predications, it is worth mentioning that Ameka & Levinson (2007) state that many languages of the world use postural verbs (most prominently *sit*, *stand*, *lie*) for the expression of locative predication so that the class of verbs possibly occurring in locative predication must be widened. Regardless of whether one subsumes postural verbs under the locational strategy or makes up a separate "postural" strategy, the general assumptions of Stassen (1997) still hold and need not be revised.

Regarding existential predication, McNally (2016) and Creissels (2019) provide the most systematic proposals of a typology. McNally (2016: 212–213) does not assume a common semantic structure of locative and existential predications and clusters the language-specific realisations of existential predication independently from locative predication. Creissels (2019: 41), in turn, states that "inverse-locational predication [encodes] the same prototypical figure-ground relationships, but with the marked perspectivization 'ground > figure'", implying that existential predication is the marked version of locative predication. In the first step of his typology, Creissels (2019: 55–57, 60–64) distinguishes languages which exhibit a designated morphosyntactic construction for the expression of existential predications (e.g. English *there is*) from languages which express existential predications via word order permutations or merely via the context. In what follows, he develops a detailed typology for languages of the former type. As valuable as Creissels' (2019) approach is, it leaves the question open of how to deal with instances of existential takeover,

[5] An anonymous reviewer correctly points out that accepting the term and process of existential takeover presupposes accepting (a) existential predication as a functional domain separate from locative predication and (b) an existential strategy (e.g., the application of semantically existential items) as the prototypical coding strategy in existential predication. Fair enough, neither of these axioms can finally be proven in this paper, but I still think that existential predication is functionally to be separated from locative predication (see Section 2.1), and it is at least not far-fetched to account for the usage of existential items as their prototypical coding strategy. Whether or not this holds cross-linguistically, can, however, hardly been answered in this paper and remains a question for further research.

as shown in (2) and described in detail in Section 4. Should a given language be classified as a "share" language, exhibiting no dedicated existential predication structure, which is counterintuitive since an existential strategy is used? Or should this language be classified as a "split" language, exhibiting a dedicated existential predication structure, which is likewise counterintuitive given that the disambiguation of locative and existential predication is only provided via word order and/or the context? Section 5 deals in more detail with these questions and resulting typological implications.

Finally, Veselinova (2013) and Veselinova & Hamari (2022) provide a comprehensive account of the expression of negative existential predication; however, the perspective chosen is on the expression of negation rather than on the expression of existential predication itself. Still, it provides essential insights, which I take up in Sections 3.3, 3.4.2 and 4 when dealing with negative existentials appearing in locative predication.

As an interim conclusion, it must be stated that to date – regardless of the extensive existing literature and many valuable approaches – there is no cross-linguistically applicable typology of locative and existential predication which recognises all necessary aspects.

3. Locative and existential predication in Siberian languages

3.1. Languages and data

"Siberian languages" is used here as a geographically motivated umbrella term for the roughly 40 languages spoken in Siberia, that is, east of the Ural Mountains in the Russian Federation. Most Siberian languages are severely endangered and at the edge of extinction (Vajda 2009: 425–428). Whereas the Siberian languages belong to different language families (Uralic, Turkic, Tungusic, Mongolic, Yeniseian, Chukotko-Kamchatkan, Eskimo-Aleut) or are linguistic isolates, many of them share several typological features, e.g. the following (see Anderson 2006 and Vajda 2009):

- rather simple vowel systems
- vowel harmony
- suffixal agglutination
- elaborate case systems with many local cases
- dependent-marking structures
- postpositions
- basic SOV word order

- word order permutations used for pragmatic purposes
- clausal subordination with nominalised verb forms

As for the expression of locative and existential predication, the most important features are the widespread possibility of locative case marking of the ground element, the basic SOV word order and the pragmatically driven word order permutations.

The paper at hand does not aim to investigate all Siberian languages but focuses on fourteen of them, spoken, as a tendency, in Western and Central Siberia. Table 1 lists the languages, their genetic affiliation, and the estimated number of speakers according to the last Russian census in 2020[6]. Additionally, it lists the sources from which I took the relevant language data.

LANGUAGE	FAMILY, GENUS	SPEAKERS	SOURCES
Khanty (kca)[7]	Uralic, Ob-Ugric	9,230	- Ob-Ugric Database (Kazym, Yugan and Surgut Khanty) - Steinitz (1975, 1989) (Sherkaly and Synja Khanty) - Annotated folklore and daily prose texts in the languages of the Ob-Yenisei linguistic area (AnnTObY) (Filchenko et al. 2010–2021) (Vasyugan Khanty)
Mansi (mns)	Uralic, Ob-Ugric	1,346	- Ob-Ugric Database (Northern and Western Mansi) - Munkácsi (1892, 1893) (Tavda Mansi)
Nenets (yrk)	Uralic, Samoyedic	24,487	INEL Nenets Corpus (both Tundra and Forest Nenets)
Forest Enets (enf)	Uralic, Samoyedic	97[8]	INEL Enets Corpus
Nganasan (nio)	Uralic, Samoyedic	300	INEL Nganasan Corpus
Selkup (sel)	Uralic, Samoyedic	975	INEL Selkup Corpus
Kamas (xas)	Uralic, Samoyedic	extinct	INEL Kamas Corpus
Dolgan (dlg)	Turkic, Northeastern	4,836	INEL Dolgan Corpus

[6] https://rosstat.gov.ru/vpn/2020 (Accessed March 21, 2024).

[7] Fair enough, as pointed out by an anonymous reviewer, Khanty is rather an umbrella term for several Khanty languages. Still, for the topic under discussion here, all Khanty varieties appear to behave similarly, so they can be dealt with together in this paper.

[8] Note that the Russian census does not differentiate Forest and Tundra Enets. Since Forest Enets is the less moribund Enets variety, it can be safely assumed that the majority of the people declaring to speak Enets are indeed speakers of Forest Enets.

LANGUAGE	FAMILY, GENUS	SPEAKERS	SOURCES
Sakha (sah)	Turkic, Northeastern	377,722	- Alekseev (1995) - Emel'janov & Smirnov (2008) - YRCSC (Yakut-Russian Code-Switching Corpus)
Chulym Turkic (clw)	Turkic, Northeastern	32	- Annotated folklore and daily prose texts in the languages of the Ob-Yenisei linguistic area (AnnTObY) (Filchenko et al. 2010–2021) - ELAR Melets Chulym collection (Filchenko 2016–2019)
Evenki (evn)	Tungusic, Northern	5,831	INEL Evenki Corpus
Even (eve)	Tungusic, Northern	5,304	- DOBES collection "Even" (Aralova et al. 2007–2023) - Sotavalta (1978)
Ket (ket)	Yeniseian	61	- Annotated folklore and daily prose texts in the languages of the Ob-Yenisei linguistic area (AnnTObY) (Filchenko et al. 2010–2021) - *Siberian Lang* database - Dul'zon (1966, 1971) - Kotorova & Porotova (2001)
Yugh (yug)	Yeniseian	extinct	- Dul'zon (1971) - Werner (1997)

Table 1: Languages and data.

Wherever possible, I used electronically searchable language corpora; otherwise, the data come from previously published text collections. In either case, it is essential to mention that the data come from coherent texts and, thus, discourses, so they have linguistic context and can be analysed for discourse-pragmatic features. The data are collected and annotated for several semantic and pragmatic features in the XML-based EXMARaLDA[9] format; afterwards, they are coded in an SPSS database[10] that allows statistical analyses and significance tests. As for analysing the data, it is important to note that the interpretation of the data has two major sources: first, the translations in the corpora are chiefly used for understanding the propositional content of an utterance in question. Second, and more importantly, the interpretation of the reading

[9] https://exmaralda.org/en/, (Accessed on March 21, 2024).

[10] https://www.ibm.com/spss, (Accessed on March 21, 2024).

(locative vs existential) is drawn from the linguistic context of the utterance in question, which is why it is so important to analyse data from coherent texts.

3.2. Affirmative clauses

As a rule, locative and existential clauses consist of three elements in the analysed languages. The figure element is coded as the unmarked subject of the clause, the ground element is a nominal, an adpositional phrase or an adverb marked for location, and the linking element provides a syntactic connection of the former two. Since both locative and existential predications inherently express a figure-ground relation, it is the linking element that cross-linguistically shows the most variation relevant for a typology of locative and existential predication.

To avoid confusion in what follows, I briefly introduce here how I define the coding strategies applied. “Zero copula” means that there is no lexical linking element in the clause. The figure referent can be indexed via person-number suffixes at the ground element, but figure and ground can also merely be juxtaposed. “Copula” means that a semantically empty copula verb, also appearing in nominal and adjectival predication, functions as the linking element. “Semi-copula” designates linking elements that are not entirely bleached, but their original meaning is still transparent. These are locative verbs like ‘be located’ or ‘be placed’, typically being restricted to locative predication and not appearing in nominal and adjectival predication, as well as postural verbs, originally describing a body posture, most prominently ‘sit’, ‘stand’, ‘lie’. “Existential” designates items that have existential semantics, which can be shown by (a) existential usages outside existential predications (e.g. nominal meanings such as ‘existence’, ‘absence’ or ‘lack’), (b) their appearance in generic existentials and (c) prototypical grammaticalisation patterns as described in Creissels (2019). Additionally, it should be noted that I structure the description according to the coding strategies not according to languages. Therefore, language-internal variation is not covered optimally, but for the sake of this paper, this can be regarded as secondary. If variation is relevant for the topic under discussion, I surely point to it.

In locative clauses, the linking element can be either a zero copula, an overt copula or a semi-copula, the latter including any locative or postural verbs. In the case of a zero copula, the figure can be cross-referred to at the ground element (9; Stassen’s (1997) *verbal strategy*), or there is no overt connecting element altogether (10; Stassen’s (1997) *nominal strategy*). The former pattern occurs systematically in Ket

and Yugh but is also present as a minor strategy in Nganasan, Dolgan and Sakha. The latter pattern is widespread in Khanty and Chulym Turkic but is occasionally also attested in all other languages. It must be noted that Russian, as the dominating contact language, also exhibits a zero-copula pattern in present-tense locative clauses (Paducheva 2008: 148), so contact-induced changes cannot be excluded.

(9) Yugh (Yeniseian; Werner 1997: 287)
xeb-ś, *ad* *uk* *fɬ·l'-iŋ-gej-**di**²*.
bear-VOC 1SG 2SG.GEN large.intestine-PL-LOC-**1SG**
'Bear, I am in your intestines.'

(10) Vasyugan Khanty (Uralic, Ob-Ugric; Filchenko et al. 2017: 33)
wajaɣ *jiɣi* *jor-nə.*
animal river middle-LOC
'The animal is in the middle of the river.'

Locative clauses containing a semantically empty copula are most frequent and widespread in Mansi, Nganasan, Selkup, Kamas, Chulym Turkic, Evenki and Even (11). Disregarding possible diachronic evolutions, I classify them as following Stassen's (1997) *nominal strategy* since, synchronically, no locative semantics of the used copulas can be singled out. Finally, Nenets and Enets use locative verbs in locative predication (12), Khanty and Mansi exhibit postural verbs (13), and Dolgan and Sakha show the existential nominal *baːr* 'exist(ing)' (see example (2) in the introductory section), which I discuss in detail in Section 3.4 and 4.

(11) Northern Evenki (Tungusic, Northern; INEL Evenki Corpus: YUK_2007_PoorPeople3_nar.037)
Oːriktə *jes'o* *N'əkəŋdə-du* ***bi-s'o-n.***
Orikte still.R Ekonda-DAT/LOC **be-PST-3SG**
'Orikte was still in Ekonda.'

(12) Forest Nenets (Uralic, Samoyedic; INEL Nenets Corpus: ALY_200206_Life_nar.003)
Šoɬ'a-j *m'a-kna* ***me-štu-t.***
Sholi-POSS1SG tent-LOC.SG **be.there-HAB-1SG**
'I was in Sholi's tent.'

(13) Sherkaly Khanty (Uralic, Ob-Ugric; Steinitz 1975: 299–300)

moχa	*taj-əm*	*topas-ŋət*	*χɔ̄t-ŋət*	*iśə*
before	have-PTCP.PST	storage-DU	house-DU	same
wot-et-na	***ɔ̄məs-t-aŋn̥.***			
place-POSS3SG-LOC	**sit-PRS-3DU**			

'The storage and the house, which he had before, are [lit. sit] at the same place.'

Existential clauses either exhibit the same morphosyntactic structure as locative clauses (Khanty, Mansi, Selkup, Kamas, Evenki, Even), or they contain an existential predicator according to Creissels' (2019) typology (Nganasan, Enets, Nenets, Dolgan, Sakha, Chulym Turkic, Ket, Yugh). In the former case, the disambiguation is guaranteed via word order permutations. For example, the Evenki locative clause (11) above shows the word order "figure – ground – copula", whereas an existential clause (14) shows the word order "ground – figure – copula". Apart from the word order permutation, there is thus no formal difference.

(14) Northern Evenki (Tungusic, Northern; INEL Evenki Corpus: BTV_20190815_ShamanNyokcho_nar.020)

utolə	*Hantajka-du*	*kətəː*	*haman'-il*	*bi-ŋki-tin.*
earlier	Khantayka-DAT/LOC	many	shaman-PL	be-PST.DIST-3PL

'Earlier, there were many shamans in Khantayka.'

In the languages exhibiting a dedicated existential pattern, the linking element may either be an existential verb (Nganasan, Enets, Nenets) (15), an existential nominal (Dolgan, Sakha, Chulym Turkic) (16) or an existential particle (Ket, Yugh) (17).[11] Though the word class of a relevant item is not immediately relevant for typologising the pattern as such, it is essential regarding its diachronic sources and assessing its initial semantics, which are discussed in detail in Section 3.4.

[11] The word-class membership of the items is derived from inflectional categories being attached: TAME morphology in the case of verbs, case and number morphology in the case of nominals and no such morphology in the case of particles. Fair enough, the given examples do not prove that the Turkic items are nominals, as opposed to the Yeniseian particles; however, relying on Johanson (2021: 817) and Georg (2007: 314), this seems to be clearly the case.

(15) Nganasan (Uralic, Samoyedic; INEL Nganasan Corpus: ChNS_080214_Wandering_nar.023)

Təndə	*ŋil'ə-mənu*	*bɨʔ*	*bɨʔ*	***təi-s'ütə.***
that.GEN.SG	bottom-PROL.SG	water	water	**EX-FUT.3SG**

'There will be water under it.'

(16) Chulym (Turkic, Northeastern; Filchenko 2016–2019: TamochevaVA_TamochevGG_05Aug2015_Self_Interview_00043-299)

üs-tä,	*üs-tä*	*palïx*	***par.***
Chulym-LOC	Chulym-LOC	fish	**EX**

'There is fish in [the river] Chulym.'

(17) Southern Ket (Yeniseian; Kotorova & Porotova 2001: 52)

[…]	*ovet-diŋt*	*nan'*	*kan*	***usaŋ.***
	lunch.R-ADE	bread	OPT	**EX**

'[In the morning, I place the dough, I prepare it for lunch,] so there is bread for lunch.'

As for the relationship of affirmative locative and existential predications, the languages under investigation can thus be grouped as follows: [12]

1) The language has one single "non-existential" morphosyntactic structure used in locative and existential predications, the disambiguation being established via word order changes: Khanty, Mansi, Selkup, Kamas, Evenki, Even.
2) The language has different morphosyntactic structures in locative and existential predications; word order changes may additionally point to a locative and existential reading, respectively: Nganasan, Enets, Nenets, Chulym, Ket, Yugh.

[12] Note that the following generalisations hold only for affirmative present tense, indicative mood. The picture becomes more intricate when adding tense or other verbal categories as parameters. Since, however, the paper at hand does not aim at a complete description of locative and existential predication patterns in the investigated languages, this can be left aside here. The following argumentation holds also, if one language is to be classified differently in other tenses, moods or the like. As for negation, see below.

3) The language has one single "existential" morphosyntactic structure used in locative and existential predications, the disambiguation being established via word order changes: Dolgan, Sakha.

The first two groups match Creissels' (2019) division of "share" and "split" languages exactly. In his terminology, the first group correlates to languages exhibiting a *general-locational predication*, disambiguated in the given context, whereas the second group of languages exhibit an *inverse-locational predication* formed by existential predicators, opposed to *plain-locational predication.* The third group, however, rather correlates to languages in which the *inverse-locational predication* loses its marked status and is reanalysed as a *general-locational predication* (Creissels 2019: 61). In the terminology applied here, the existential predication pattern is generalised and taken over to locative predication, thus exhibiting a strategy takeover in Stassen's (1997) sense, which I label *existential takeover*. Since this process seemingly contradicts frequent assumptions on the functional (un)markedness of locative and existential predication, I discuss it amply in Section 4 from this perspective.

3.3. Negative clauses

As for the negation of locative and existential predications in the analysed Siberian languages, one clear tendency is observable: "non-existential" structures in locative predications are given up for the benefit of "existential" structures. In all fourteen languages, negative existential items are at least partially constitutive for negated locative and existential predications, as exemplified by Khanty (18–19), Selkup (20–21) and Evenki (22–23). In either example, the first clause shows a locative predication and the second clause an existential predication. Note that *u-* 'be' in Khanty (18) is necessary for the expression of tense since the past tense marker *-s* must not be attached to the negative existential particle. Consequently, *u-* is not used as a copula element to connect the subject and predicate but rather as an auxiliary.

(18) Sherkaly Khanty (Uralic, Ob-Ugric; Steinitz 1989: 168)

śeman	*jŏtn̥*	***ăntɔm***	*u-s.*
Semyon	at.home	**NEG.EX**	be-PST.3SG

'Semyon was not at home.'

(19) Vasyugan Khanty (Uralic, Ob-Ugric; Filchenko et al. 2020: 56)
jiɣi-nə *muɣi* ***əntim.***
river-LOC crucian **NEG.EX**
'There are no crucians in the river.'

(20) Southern Selkup (Uralic, Samoyedic; INEL Selkup Corpus: SUF_1967_DaughterAndRobbers_flk.242)
Mi *ta-nan* ***t'aŋ-sa-ut.***
1PL.PRO 2SG.PRO-ADE **NEG.EX-PST1PL**
'We were not at your place.'

(21) Southern Selkup (Uralic, Samoyedic; INEL Selkup Corpus: SEV_1967_ThreeSisters_flk.018)
n'äj *maǯi-gu* *mat-qɨt* *paɨ* ***t'äŋg-wa.***
bread stab-INF tent-LOC knife **NEG.EX-CO.3SG**
'There is no fish in the fishing net.'

(22) Northern Evenki (Tungusic, Northern; INEL Evenki Corpus: BTV_20190820_Pankagir_nar.011)
ami-w-ka ***a:sin*** *bi-so-n* *moha-du.*
father-POSS1SG-EMPH **NEG.EX** be-PST-3SG taiga-DAT/LOC
'My father was not in the taiga, [but in the settlement].'

(23) Southern Evenki (Tungusic, Northern; INEL Evenki Corpus: BaN_1930_FoxAndWolverine_flk.039)
d'u:-du-wị ***a:čịn*** *d'əptilə-l.*
house-DAT/LOC-RFL.POSS.SG **NEG.EX** food-PL
'There is no food at home.'

As noted by an anonymous reviewer and spelt out by Panova & Liljegren (forthcoming), among others, negative locative clauses are hard to discriminate against negative existential clauses, because a negative locative clause presupposes the existence of the figure – since it is the perspectival centre – but denies its presence in the given location. Therefore, the negation in locative clauses must not scope over the whole clause, but only over the ground element, yielding contrastive focus

structures (X is not at Y, [but at Z]). Such contexts are rare in natural speech, and the analysed material contains less than fifty clear instances of negative locative clauses altogether. The examples shown above, however, fulfil this criterion. In (18), the figure referent (Semyon) is introduced in the left context, but not the ground referent (at home). Thus, the latter is not presupposed and cannot be the perspectival centre of the utterance, so the example cannot be analysed as an existential clause. In (20) and (22), the speaker talks about the places of being of the figure referents, so the perspectival centre of the utterances is again the figure referent.

For the sake of completeness, however, it should be noted that "non-existential" strategies are also used, triggered by various morphosyntactic parameters. For example, Chulym Turkic shows a split between TAME-unmarked and TAME-marked forms (see Däbritz 2024 for details). In the former, the negative existential *čok* ~ *čoɣul* functions as the linking element in both locative and existential predications (24–25), whereas it is the copula *pol-* 'be(come)' in the past tense (26–27). Again, both (24) and (26) are to be classified as locative clauses since the figure is the perspectival centre of the utterance, which can be derived from the left context in the source material. Additionally, in (26), the speaker lists the places where they have been or not, evoking a contrastive list reading.

(24) Chulym (Turkic, Northeastern; Filchenko et al. 2010: 297)
čilyə-zə *minda* ***čoɣul.***
horse-POSS3SG here **NEG.EX**
'The horse is not here.'

(25) Chulym (Turkic, Northeastern; Filchenko 2016–2019: TamochevaVA_05Aug2015_Self_Interview_00042_1-27)
Pasečnaj-da *škol* ***čoɣul.***
Pasechnoe-LOC school **NEG.EX**
'There is no school in Pasechnoe.'

(26) Chulym (Turkic, Northeastern; Filchenko 2016–2019: Kondiyakov_Gabov_July2016_Meeting-1.44)
nu, nu, män *Töɣöldet-tä* *pir ras-ta* ***pol-v-a-m.***
well well 1SG Teguldet-LOC one time-LOC **be-NEG-PST-1SG**
'Well, I wasn't a single time in Teguldet.'

(27) Chulym (Turkic, Northeastern; Filchenko 2016–2019: KondiyakovAF_06Aug2015_Interview_00024_1-55)

a	*an-da*	*nerva-lor-u*	***pol-v-an.***
and	that-LOC	nerve-PL-POSS3SG	**be-NEG-PST.3SG**

'And there were no nerves there [= under the teeth]. '

Given the observed structures, one might wonder whether the negative existential items included are indeed existentials. In Section 3.4, this question is targeted, and it is shown that both synchronic and diachronic arguments favour treating them as true existentials regarding their lexical source. Given this, Section 4 analyses also the existential takeover in negative locative predications from the perspective of markedness and related issues.

3.4. Sources of existential items

3.4.1. Existential nominals in Dolgan and Sakha

This section deals with the role of existential nominals in locative and existential predication in the Northern Siberian Turkic languages Dolgan and Sakha. I will focus on the affirmative existential nominal *baːr* 'exist(ing)' in this section, whereas its negative counterparts *huo͡k* (Dolgan) and *suo͡χ* (Sakha), respectively, are more closely analysed in Section 3.4.2. As noted already in the introduction, Dolgan and Sakha express affirmative locative and existential predications employing the affirmative existential nominal *baːr* 'exist(ing)'. Examples (28–29) show locative clauses in these languages, and examples (30–31) show existential clauses. As can be seen, only the context and word order differentiate the locative from the existential reading. (28) is the answer to the question "where are you", so the ground element is focused and, thus, evokes a locative reading. In (29), several people are playing monopoly, and one of them states that another must not throw the dice because he is in jail, which again evokes a locative reading. In (30), in turn, the speaker hands a pocket to her son and now explains what is inside. Thus, the figure element is necessarily included in the focus domain. (31) works similarly since the speaker tells what was there on the way.

(28) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: PoS_PrG_1964_Lyybyra_flk.076)
D'ie̯ *ih-i-ger* ***ba:r-bin.***
house inside-POSS3SG-DAT/LOC **EX-1SG**
'I am in the house.'

(29) Sakha (Turkic, Northeastern; YRCSC, own glossing)
Xaji:-ga ***ba:r.***
jail-DAT/LOC **EX**
'He is in jail.'

(30) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: ErSV_1964_WarBirdsAnimals_flk.442)
Bu *karma:ŋ-na-r* *mö:čük* ***ba:r,*** *hüter-eje-gin.*
this pocket-POSS2SG-DAT/LOC ball **EX** lose-ADM-2SG
'There is a ball in your pocket, do not lose it.'

(31) Sakha (Turkic, Northeastern; Emel'yanov & Smirnov 2008: 313)
[…] *a:ra* *kie̯ŋ* *nali:* *u:* ***ba:r*** *ebit.*
on.the.way broad spilling water **EX** EVID
'[When he was going,] there appeared to be broad, spilling water on the way.'

For the sake of completeness, it should be mentioned that in other tenses and moods than present indicative, the existential nominal *ba:r* is supported by a form of the copula/auxiliary verbs *e-* 'be' and *bu̯ol-* 'be(come)'. Examples (32–33) and (34–35) illustrate this.

(32) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: KiES_KiLS_2009_Life_nar.KiES.001)
D'e *Korgo:-go* ***ba:r*** ***e-ti-bit.***
well Korgo-DAT/LOC **EX** **be-PST1-1PL**
'Well, we were in Korgo.'

(33) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: SiAN_2008_LifeInTundra_nar.SiAN.103)
Urut *kerget-ter-im* ***baːr*** ***er-dek-terine*** […].
earlier parents-PL-POSS1SG **EX** **be-COND-3PL**
'Earlier, when my parents were still there, [they taught me].'

(34) Literary Sakha (Turkic, Northeastern; online data[13], own glossing)
Min *ikki-s* *bölöχ-χö* ***baːr*** ***e-ti-m.***
1SG two-ORD group-DAT/LOC **EX** **be-PST1-1SG**
'I was in the second group.'

(35) Sakha (Turkic, Northeastern; YRCSC, own glossing)
Ikki *štuka* ***baːr*** ***bu͡ol-u͡oγ-a.***
two piece.GEN.R **EX** **become-FUT-3SG**
'There should be two pieces.'

Additionally, both locative and existential clauses can lack the existential nominal *baːr*, as displayed by the Dolgan examples (36–37).

(36) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: SuON_KuNS_19990303_HardLife_conv.SuON.253)
Patap-ka ***e-ti-bit.***
Potapovo-DAT/LOC **be-PST1-1PL**
'We were in Potapovo.'

(37) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: PoPD_KuNS_2004_Life_conv.PoPD.042)
Avaːm-ŋa *onno* *taba-lar* *agaj* ***e-ti-lere.***
Ust.Avam-DAT/LOC there reindeer-PL only **be-PST1-3PL**
'There were only reindeer there in Ust-Avam.'

[13] https://www.s-vfu.ru/universitet/rukovodstvo-i-struktura/instituty/iyikn/news/detail.php?SECTION_ID=&ELEMENT_ID=43700, (Accessed on June 23, 2023).

Thus, the copula/auxiliary verb (32–35) and the omission of the existential nominal *baːr* (36–37) concern both locative and existential predications. From a statistical point of view, the existential nominal is more frequent in existential than in locative clauses in either language (EX: 55.9% vs. LOC: 40% in Dolgan, and EX: 73.4% vs. LOC: 42.2% in Sakha). However, since (zero-)copula structures regularly appear in both domains, too, neither the occurrence of the existential item nor its lack can disambiguate locative and existential readings. Therefore, I do not discuss this issue here further.

Applying Creissels' (2019) typology, Dolgan and Sakha belong to the group of "share" languages since the morphosyntax of locative and existential predications is identical. However, the existential nominal *baːr* 'exist(ing)', as well as its negative counterpart *huo͡k* ~ *suo͡χ*, has precise existential semantics, which can be proven both synchronically and diachronically.

First, either item is undoubtedly nominal from a morphological point of view (Ubrjatova et al. 1982: 440; Däbritz 2022: 69–70). This can be shown by lexicalised light verb constructions such as, e.g. Sakha *baːr gin-* 'have available; have in stock' (lit. 'make existent') and *suo͡χ gin-* 'liquidate; defeat' (lit. 'make non-existent') (Ubrjatova et al. 1982: 112, 404), in which the existential syntactically occupies the direct object position. Also, on a synchronic level, either item can occur in argument and adjunct positions without further derivation, as exemplified by (38) and (39).

(38) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: BeES_2010_HidePreparation_nar.030)

Tuo͡k	*kuhagan*	***baːr-in***	*bari-tin*	*iti*
what	bad	**EX-POSS3SG.ACC**	all-POSS3SG.ACC	that

il-atta-n *ih-el-ler.*
take-MULT-CVB.SEQ go.AUX-PRS-3PL
'They take away everything bad that is there.'

(39) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: PoPD_KuNS_2004_Life_conv.PoPD.106)

Urut	*otto*	*karči*	***huo͡g-u-ttan***	[…].
earlier	then	money	**NEG.EX-POSS3SG-ABL**	

'Earlier because of the lack of money, [if you want to do something, well…].'

Returning to the existential nominal *baːr* ‘exist(ing)’ in locative and existential predication, its appearance in existential predication is entirely expected given its existential semantics. Consequently, its appearance in locative predication seems less expected from a language-internal perspective of Sakha and Dolgan, and it appears to be a secondary usage. Additionally, it frequently appears in generic existential clauses (Däbritz 2024: 86, 100), which also underlines its existential semantics in the given context.

The secondariness of the appearance of Sakha and Dolgan *baːr* ‘exist(ing)’ in locative predications can also be underpinned from a historical-comparative perspective. Both *baːr* (affirmative) and *huo͡k* ~ *suo͡χ* (negative) are inherited from earlier stages of Turkic, which can easily be shown by their cognates in various Turkic languages of different branches, e.g. Turkish (tur) *var* and *yok*, Bashkir (bak) *bar* and *juq*, Chuvash (chv) *por* and *śuk* ~ *śok*, Kirghiz (kir) *bar* ~ *ʤok* (Karakoç 2009: 218; Miščenko 2017: 111–112; Baranova et al. 2021: 11, 20; Johanson 2021: 484, 817). In most Turkic languages, however, the affirmative existential nominal is restricted to existential and possessive predications. In contrast, a (zero) copula or person-number agreement suffixes are the linking element in locative predication. As a case in point, Bashkir uses *bar* ‘exist(ing)’ in the former types (40a–b) but a (zero) copula in the latter (40c).

(40) Bashkir (Turkic, Kipchak; Miščenko 2017: 121, glossing adapted)

a.	*Beð-ðeŋ*	*awəl-da*	*magazin*	***bar.***	
	1PL-GEN	village-LOC	shop	EX	
	‘There is a shop in our village.’				EXISTENTIAL
b.	*Mineŋ*	*mašina-m*	***bar.***		
	1SG.GEN	car-POSS1SG	EX		
	‘I have a car.’				POSSESSIVE
c.	*Mineŋ*	*kitab-əm*	*öθtäl-dä.*		
	1SG.GEN	book-POSS1SG	table-LOC		
	‘My book is on the table.’				LOCATIVE

In this context, it is worth noting that several Turkic languages spoken in Iran, e.g. Khorasan (kmz) and Khalaj (klj), pattern like Dolgan and Sakha. They likewise allow affirmative existential nominals in locative predications, as demonstrated by examples (41–42).

(41) Khorasan (Turkic, Oghuz; Karakoç 2009: 219, glossing adapted)
Ev-dä ***baːʼr-am.***
house-LOC **EX-1SG**
'I am at home.'

(42) Khalaj (Turkic, Khalaj; Karakoç 2009: 219, glossing adapted)
Iːraːn-ča ***vaːʼr-am.***
Iran-LOC **EX-1SG**
'I am in Iran.'

Karakoç (2009: 221–222) demonstrates that the usage of the existential nominal in locative predications, i.e. the existential takeover, in Turkic languages of Iran is a contact-induced pattern copied from Indo-Iranian. Not going into details of Indo-Iranian existential predications, this assumption seems plausible since the relevant Turkic varieties spoken in Iran belong to different branches of Turkic, Khalaj being argued to be a branch of its own (Johanson 2021: 21–23, 91–92). Given the general Turkic picture discussed above, the Khorasan and Khalaj patterns can hardly be traced back to a common origin. A parallel development without an external motivation is principally possible, but given the surrounding Indo-Iranian languages, which are dominant in either case, the contact-induced explanation is more solid.

Hence, both Dolgan/Sakha and the Turkic varieties spoken in Iran exhibit patterns of existential takeover, which developed independently. In the latter case, the takeover is most probably contact-induced, whereas this can be excluded in the former case since the surrounding languages of Dolgan and Sakha do not exhibit it either. Maybe, Mongolic languages might have influenced Pre-Dolgan-Sakha since they also show shared patterns of locative and existential predications (Janhunen 2003: 26–27). Still, this cannot be proven. Be the takeover process synchronic or diachronic, one can say that Dolgan and Sakha – as well as the Turkic varieties in Iran – exhibit existential nominals, constitutive for both existential and locative predications. In either case, synchronic and diachronic data prove that their existential reading is primary, so one must conclude that the existential nominals have spread to locative predication.

3.4.2. Negative existentials in Siberian languages

Before discussing negative existentials, some general properties of negated locative and existential predications must be clarified. Semantically, the negative sentence *the book is not on the table* can be relatively easily decomposed into its affirmative counterpart and a negative operator.[14] Pragmatically, however, the given utterance functions differently from its affirmative counterpart. It has long been noted (Givón 1978, Tottie 1991, Miestamo 2005, among others) that negative utterances are only felicitous in specific discourse contexts, which all implicitly or explicitly presuppose the affirmative counterpart of the utterance in question. Hence, *the book is not on the table* is pragmatically adequate only if the context somehow suggests that the book principally might be on the table (Miestamo 2005: 197–198). Regarding information structure, the relevant contexts often correlate with verum focus (43) or contrastive focus (44) constructions.

(43) A: *Can I have a look into the book you recently bought?*
B: *Sure, take it. It is on the table over there.*
A (looking for the book): *No, it is NOT on the table.*

(44) A: *Please, go and get me my diary from my desk.*
B (returning): *Here you are. However, it was not on the DESK, but on the SHELF.*

Given these pragmatic constraints, one should expect that negative locative predications are significantly less frequent than their affirmative counterparts. Indeed, this expectation holds: In the analysed material, there are 1,059 affirmative locative clauses but only 49 negative locative clauses (95.6% vs 4.4%). As for existential predications, the same tendencies can be observed, though the share of negative existential clauses is significantly higher in the analysed material: there are 2,625 affirmative existential clauses and 1,164 negative existential clauses (69.3% vs 30.7%). This can be explained by the central discourse function – the (re-)introduction of a discourse referent (see Section 2.1, Hengeveld 1992, Dryer 2007) – of existential

[14] Note that *decomposition* and *operator* must not be understood in formal semantic terms here, but only to illustrate the problem. Interestingly, some languages, e.g. Vietnamese, reflect this decomposed semantics in negative non-verbal predications, which literally can be translated “it is not true that…” (Eriksen 2011: 280; Veselinova & Hamari 2022: 43).

predications that is supposedly more compatible with negation than asserting a location to a given referent: When talking about a situation in general, it is often an equally adequate information that something is absent. In turn, this is more problematic when "zooming" on the absent referent, since its existence per se must not negated, but only its episodic presence, because it is necessarily presupposed and, thus, existing. Consequently, the amount of data to be analysed for negative existential takeover is quantitatively quite restricted. Still, the available data show evident patterns.

As shown in Section 3.3, all languages under discussion here use negative existential items to express negative existential predication, Kamas (45) again illustrating that. Additionally, all languages exhibit the same negative existential item in negative locative clauses, as displayed by Kamas (46).

(45) Kamas (Uralic, Samoyedic; INEL Kamas Corpus: PKZ_1964_SU0207.PKZ.094)
Ma?-nan *sazən* ***naga.***
tent-LAT/LOC.POSS2SG paper **NEG.EX**
'There is no paper at your home.'

(46) Kamas (Uralic, Samoyedic; INEL Kamas Corpus: PKZ_196X_AngryLady_flk.044)
Da *tăn* *gijen-də* *i* ***nago-bia-l*** […].
and.R 2SG where-INDEF and.R **NEG.EX-PST-2SG**
'But you haven't been anywhere, [you lived here].'

Taking the negative existential semantics of inter alia Kamas *naga* ~ *nago-* for granted, Kamas – as well as the other thirteen languages of the sample – appear to exhibit a systematic existential takeover in negative locative predications. However, it must be shown again that the negative existential items indeed have existential semantics. According to Veselinova (2013: 139) and Veselinova & Hamari (2022: 34–41), negative existential items are hardly mere negators of affirmative existentials but rather replace affirmative existentials; following Eriksen (2011: 281–283), they represent a "direct negation avoidance strategy". Thus, their semantics include both negation and 'existence', which leads to a reading of 'absence' (ibid.). Consequently, Veselinova & Hamari (2022) argue against the compositional semantics of negative

existentials. Going a step further, they take this assumption as a reason for the fact that negative existentials diachronically often trace back to sources like ‘lack’, ‘absent’ or ‘empty’ but are rarely formal compositions of a negative item and an affirmative existential (Veselinova & Hamari 2022: 38–39). This observation is also relevant to the topic under discussion here. Given a negative existential item with the initial meaning ‘lack’, which appears in negative locative and existential predications in language X, it can hardly be argued that locative predication is ontologically primary against existential predication in this language. Therefore, I analyse the negative existential items of the languages under discussion here from a diachronic perspective in what follows.

First, it can be stated that the data underpin Veselinova & Hamari’s (2022) claim that negative existentials often trace back to full lexical items indicating absence. Only the Yeniseian languages Ket and Yugh display a univerbation pattern, namely Ket *bənsaŋ* and Yugh *bəše*, which originate in the combination of the negative particle *bən* and the affirmative existentials *useŋ* and *uše*, respectively (Werner 1997: 215; Georg 2007: 314).

The Ob-Ugric languages Khanty and Mansi show the following negative existential items: Northern Khanty *antəːm* ~ *ăntɔm* ~ *antum*, Eastern Khanty *əntem* ~ *əntim*, Northern Mansi *atim*, Eastern Mansi *øɑ̯tʲi*, Western Mansi *oɒ̯tʲəm*, Tavda Mansi *iikəm*. As argued by Steinitz (1967: 123–124) and Veselinova (2015: 567–568), all forms trace back to a nominalisation of the Proto-Ob-Ugric negative auxiliary verb **ə-*. The nominal character of the forms can be shown by the need for a copula support item in non-present tense contexts (47). Regarding their semantics, the Northern Khanty lexicalisation *ăntɔma jĭ-* ‘die’, which literally means ‘become absent’ ~ ‘become non-existent’, neatly shows the negative existential’s semantic content (48; Steinitz 1967: 123).

(47) Eastern Khanty (Uralic, Ob-Ugric; OUDB Yugan Khanty (2010–) Corpus: Text 1615, 163)

tʲi	*puɯrnə*	*mɛŋk*	***ɛntem***	*tʲi*	***βoł.***
this	after	spirit	**NEG.EX**	so	**be.PST.3SG**

‘After this, there were no more Menks [= kind of spirit].’

(48) Northern Khanty (Uralic, Ob-Ugric; Steinitz 1967: 123)

[*χu*]	*ăntɔm-a*	*jĭ-s.*
man	NEG.EX-LAT	become-PST.3SG

‘[The man] died.’

The Samoyedic languages exhibit the following negative existential items: Nganasan *ďaŋku ~ ďaŋguj-*, Forest Enets *dʲago-*, Tundra Enets *dʲigu-*, Tundra Nenets *jəŋku-*, Forest Nenets *d'iku- ~ t'iku-*, Northern Selkup *č'ä:ŋki-*, Southern Selkup *t'äŋu-* and Kamas *naga ~ nago-*. Additionally, the extinct Samoyedic language Mator exhibits the verbal form *nagajga* (< *naga* 'NEG.EX' + *äj-* 'be' + the present tense co-affix *-ga*), which means 'there is not; there lacks' (Helimski 1997: 209, 312–313). According to Janhunen (1977: 40–41), all items can be traced back to Proto-Samoyedic **jänkV ~ jänkV-* meaning 'not, absence, missing; not be there, miss'. Apart from the usage in negative existential predications, there is only scarce evidence for initial existential semantics in the Samoyedic languages. However, the named Mator form *nagajga* can also be used to form caritive adjectives, such as, e.g. *teništa nagajga* 'stupid', literally meaning 'is without mind' (ibid.). This pattern is indeed an argument in favour of the initial existential semantics of the item since no ground element is included conceptionally. Surely, it is difficult to transfer this to the other Samoyedic languages since the latter do not exhibit similar patterns synchronically, so the Samoyedic negative existentials are not as clearly existential in their origin as could be shown for the Ob-Ugric languages. Still, an analogue interpretation is at least possible and plausible given that the diachronic source 'lack' for negative existentials is cross-linguistically widely attested (Veselinova 2013: 118–121). Another evidence for initial existential semantics can be provided by further derivations of the negative existential verb in Kamas, as displayed in (49). Here, the momentaneous derivation yields the reading 'disappear', which might be paraphrased as 'becoming absent', so the negative existential verb may also read as 'being absent'.

(49) Kamas (Uralic, Samoyedic; INEL Kamas Corpus: PKZ_196X_SU0225.241)

[…]	*i*	*sima-t*	***nago-lu?-pi.***
	and	eye-POSS3SG	**NEG.EX-MOM-PST.3SG**

'[She shot him in the eye with an arrow,] he lost his eye (lit. his eye disappeared).'

The Turkic and Tungusic languages under consideration again show more convincing evidence that the negative existential items have initial existential semantics. Dolgan *huo͡k*, Sakha *suo͡χ* and Chulym Turkic *čoːɣul* ultimately go back to the Common Turkic form **yoq* (Johanson 2021: 817). Like their affirmative counterparts, the Dolgan and Sakha forms even synchronically may function as nouns with the meaning 'lack;

absence', as discussed in Section 3.4.1 and again displayed in (50–51). So, following the same argumentation provided above, it can safely be stated that the Turkic negative existential items have initial existential semantics.

(50) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: PoPD_KuNS_2004_Life_conv.PoPD.106)

Urut	*otto*	*karči*	***huo͡g-u-ttan***	*ze*	[…].
earlier	then	money	**NEG.EX-POSS3SG-ABL**	EMPH.R	

'Earlier because of the lack of money, [if you want to do something, well… to help somebody, from what, you have no budget].'

(51) Sakha (Turkic, Northeastern; Böhtlingk 1851: 9)

[…]	*tu:s*	***suoɣ-u-ttan***	[…].
	salt	**NEG.EX-POSS3SG-ABL**	

'[These beautiful fish apparently get lost for two reasons:] The lack of salt [and because the people are used to it].'

Additionally, Dolgan exhibits the paraphrase *huo͡k buo͡l-* 'become absent' for 'die' (52), which – like in Khanty, discussed above – points to the existential semantics of *huo͡k*.

(52) Dolgan (Turkic, Northeastern; INEL Dolgan Corpus: KiES_KiLS_2009_Birth_nar.KiES.079)

"Kaja	*ogo-but*	***huo͡k***	***buo͡l-but",***	*d-i:l-ler*	*araj.*
well	child-POSS1PL	**NEG.EX**	**become-PST2.3SG**	say-PRS-3SG	just

'"Well, our child has died", they just say.'

The Tungusic languages Evenki and Even exhibit the negative existentials *a:čin* ~ *a:sin* and *ač'č'a* ~ *ač'*, respectively. They have cognates all over the Tungusic language family, although not all formal aspects are solved from a comparative point of view (Hölzl 2015: 134–135). As for the semantics of the negative existential nominals, two phenomena – which were already discussed for other languages of the sample – point to their initial existential semantics. First, they form translational equivalents of caritive adjectives, cf. the Even example (53). Second, their combination with the copula *o-* 'become' can yield the reading 'disappear' and 'die', as displayed by the Evenki examples (54–55).

(53) Even (Tungusic, Northern; Benzing 1955: 30, own glossing)
tar-al *asa-l* ***ač'*** ***hut-l-əː-səl.***
this-PL woman-PL **NEG.EX** **child-PL-PTV-PL**
'These women are childless.' ~ 'These women are without children.' ~ 'These women have no children.'

(54) Taimyr Evenki (Tungusic, Northern; INEL Evenki Corpus: NNR_190X_StrongBoy_flk.083)
[…] *taduk* ***ačin*** ***o-da-n.***
then **NEG.EX** **become-AOR-3SG**
'[So the small human said,] then he disappeared.'

(55) Northern Evenki (Tungusic, Northern; INEL Evenki Corpus: ChAD_20180923_BurbotsEvenks_flk.ChAD.010)
[…] *dəg-il,* *bəjŋ-ol* ***aːsin*** ***oː-da.***
bird-PL animal-PL **NEG.EX** **become-AOR.3PL**
'[All inhabitants of the world,] birds and animals, died.'

Putting the discussion in a nutshell, one must conclude that all negative existential items display initial existential semantics. This is well in line with Veselinova & Hamari's (2022) assumption that negative existentials are not only negators of affirmative existentials but compose negative and existential semantics. Given this, their appearance in locative predications must be analysed as secondary, which is, thus, another argument in favour of the "existential takeover"-analysis. Sections 4 and 5 discuss which implications this pattern has for the analysis and typology of locative and existential predication in general.

4. Existential takeover: markedness, frequency, complexity and salience

Starting from the information-structural and cognitive patterns described in Section 2.1, many authors implicitly or explicitly assume locative predication is unmarked and existential predication is marked. Taking this for granted and considering the functional approach discussed in Section 2.1, it is not far to seek to establish Hengeveld's (1992: 118–121) [±presentative] as the markedness exponent for

locative and existential predications. Following this approach, locative predication is unmarked ([-presentative]), and existential predication is marked ([+presentative]).

However, neither the term *markedness* nor the concept designated by it is uncontroversial, as described inter alia by Haspelmath (2006) and Bybee (2011). Opposed to the wide and often fuzzy use of the term, I narrow it here to a rather traditional reading, namely the presence or absence of a phonological or semantic feature (Haspelmath 2006: 26; Bybee 2011: 137–138, 141–142). According to Greenberg (2005[1966]: 14, 31) and Bybee (2011: 143–144), the central property associated with markedness is text frequency: marked items appear less frequently in texts than their unmarked counterparts. Furthermore, it is widely accepted that the markedness of a given linguistic structure has several corollaries that predict its linguistic behaviour, which I briefly discuss in what follows.

First, marked items tend to be formally more complex than their unmarked counterparts. This phenomenon is especially well visible in phonology since the higher complexity of an item can be measured in terms of the physical effort the speaker has to take to produce it (Greenberg 2005[1966]: 70). But also in morphosyntax, marked items are, as a tendency, more complex than their unmarked counterparts, as, e.g. English comparative and superlative forms of adjectives, cf. *big* ~ *bigger* ~ *biggest* (Bybee 2011: 143–144). These forms additionally show another tendency. If there is zero-expression in a given domain, it is the unmarked structure exhibiting it, e.g. the English positive degree of adjectives displayed above or the singular form of count nouns, e.g. English *book*-Ø opposed to plural *book-s* (Greenberg 2005[1966]: 26–27; Bybee 2011: 143). Finally, if the formal distinction of unmarked and marked items is levelled, as a tendency, the unmarked structure is generalised. E.g., in the regularisation of English past tense forms, e.g. *weep* ~ *weep-ed* instead of *weep* ~ *wept*, the less marked present-tense stem spreads to the past tense (Bybee 2011: 135) and not the other way around. Finally, Bybee (ibid.) points out that children learn unmarked before marked structures in first-language acquisition.

Applying these corollaries of markedness distinctions to locative and existential predication, we should expect the following tendencies:

1) Locative predication as the unmarked structure is more frequent in natural language than existential predication as the more marked structure.

2) The linguistic expressions of existential predication are, as a tendency, more complex than the linguistic expressions of locative predication.

3) If one of the predication types exhibits a kind of zero expression, e.g. a zero copula, it is expected to appear in locative predication.

4) If one of the correlating linguistic structures is generalised, the locative predication structure is expected to spread to existential predication and not vice versa.

5) Children are expected to acquire locative predication before existential predication.

Discussing the issues (1) – (4), I will critically assess whether a markedness-based approach to the distinction of locative and existential predication is feasible and leads to good results.

The cases of existential takeover, amply described in Section 3, challenge the described markedness-based approach to locative and existential predication with the markedness exponent [±presentative] radically. Both affirmative and negative existential items appear systematically as linking elements of locative clauses in the analysed languages from Northern Siberia. Both synchronic and diachronic data can convincingly prove the initial existential semantics of the items. So, arguing from this perspective, existential predications are by no means more marked than their locative counterparts in the analysed languages but regularly serve as the base for co-expression patterns. Fair enough, one or two counterexamples do not suffice for overturning a cross-linguistically observed correlation completely, and indeed, many languages – such as Finnish, as shown in Section 1 – generalise their locative clause structure to existential clauses. However, given the lack of a general typology of locative and existential predication (see Section 2.2), no empirically valid conclusions about the cross-linguistic frequency of locative and existential takeover, respectively, can be drawn by now. Note additionally that also Creissels (2019: 61) points to a seemingly similar case of existential takeover in Juba-Arabic (pga; Arabic-based creole spoken in Sudan). Given this, I emphasise here that further language-specific and cross-linguistic studies of these takeover patterns are highly demanded.

Besides the existential takeover patterns, the analysed language data provide another evidence relevant to the question of whether markedness plays a role in the distinction of locative and existential predication. Above, it was argued that the critical feature of unmarked linguistic structures is their high textual frequency. Consequently, assuming existential predications being marked against locative predications entails the expectation that locative predications are more

frequent in natural language than existential predications. The analysed data, however, again contradict this expectation. As Table 2 shows, existential predications are, as a rule, more frequent than locative predications; in many languages, they are twice, thrice or even four times as frequent.

	LOCATIVE				EXISTENTIAL			
	ABSOLUTE	RELATIVE	95% CI	99% CI	ABSOLUTE	RELATIVE	95% CI	99% CI
Khanty	97	24.0%	19.8% – 28.2%	18.5% – 29.5%	307	76.0%	71.8% – 80.2%	70.5% – 81.5%
Mansi	40	16.7%	12.0% – 21.4%	10.5% – 22.9%	199	83.3%	78.6% – 88.0%	77.1% – 89.5%
Nenets	61	18.4%	14.2% – 22.6%	12.9% – 23.9%	270	81.6%	77.4% – 85.8%	76.1% – 87.1%
Enets	96	33.4%	27.9% – 38.9%	26.2% – 40.6%	191	66.6%	61.1% – 72.1%	59.4% – 73.8%
Ngana-san	86	30.0%	24.7% – 35.3%	23.0% – 37.0%	201	70.0%	64.7% – 75.3%	63.0% – 77.0%
Selkup	131	26.0%	22.2% – 29.8%	21.0% – 31.0%	372	74.0%	70.2% – 77.8%	69.0% – 79.0%
Kamas	74	17.0%	13.5% – 20.5%	12.4% – 21.6%	361	83.0%	79.5% – 86.5%	78.4% – 87.6%
Sakha	44	20.9%	15.4% – 26.4%	13.7% – 28.1%	167	79.1%	73.6% – 84.6%	71.9% – 86.3%
Dolgan	181	20.9%	18.2% – 23.6%	17.3% – 24.5%	686	79.1%	76.4% – 81.8%	75.5% – 82.7%
Chulym	31	16.9%	11.5% – 22.3%	9.8% – 24.0%	152	83.1%	77.7% – 88.5%	76.0% – 90.2%
Evenki	77	19.5%	15.6% – 23.4%	14.4% – 24.6%	317	80.5%	76.6% – 84.4%	75.4% – 85.6%
Even	92	23.7%	19.5% – 27.9%	18.1% – 29.3%	297	76.3%	72.1% – 80.5%	70.7% – 81.9%
Ket	80	24.9%	20.2% – 29.6%	18.7% – 31.1%	241	75.1%	71.4% – 79.8%	68.9% – 81.3%
Yugh	18	37.5%	23.7% – 51.3%	19.3% – 55.7%	29	62.5%	48.7% – 76.3%	44.3% – 80.7%

Table 2: Number of locative and existential predications.

More technically speaking, in most datasets of the analysed languages, existential predications outnumber locative predications significantly, relying on 99%

confidence intervals. This does not hold for the Yugh dataset, whose 99% confidence interval is 36.4% around *p*, due to its small basic population. Still, also Yugh displays almost twice as many existential than locative clauses, which underlines the overall tendencies at least impressionistically; additionally, the Yugh data are statistically significant if relying on weaker 90% confidence intervals (LOC: 25.9% – 49.1%; EX: 50.9% – 74.1%).

As a disclaimer, it must be acknowledged that most datasets are biased towards folklore and other narrative texts. So, it cannot be excluded that the observed frequency patterns are symptomatic for this genre but may differ in different genres and domains of natural speech. However, the Dolgan dataset may count as a control set inasmuch the source database, the INEL Dolgan Corpus, also contains a significant amount of free conversations, the utterances included in them making up just under 25 per cent of the utterances in the whole corpus (3,221 out of 14,078). Table 3 shows that the relative number of locative predications is slightly higher in conversations than in narrative texts but not significantly from a statistical point of view when relying on a 95% confidence interval.

	LOCATIVE			EXISTENTIAL		
	ABSOLUTE	RELATIVE	95% CI	ABSOLUTE	RELATIVE	95% CI
conversations	65	24.5%	19.3% – 29.7%	200	75.5%	70.3% – 80.7%
non-conversations	116	19.3%	16.1% – 22.5%	486	80.7%	77.5% – 83.9%

Table 3: Number of locative and existential predications – Genres in Dolgan.

So, it can carefully be concluded that the genre of a text does not play a significant role regarding the text frequency of locative and existential predications. Even if assuming that the conversational data represent the "truth" better than the non-conversational data, existential clauses are still more than twice as frequent as locative clauses. Therefore, I assume that existential predications are more frequent in natural speech than locative predications. This conclusion again tackles assuming existential predications being marked opposed to locative predication.

Instead, taking a markedness opposition as such for granted, two criteria – the observed Siberian generalisation patterns and text frequency – would predict locative predications being marked and existential predications being functionally unmarked.

A third criterion, the degree of complexity of the correlating linguistic structures, in turn, still points towards deeming existential predications marked. As a rule, existential clauses are more complex than their locative counterparts on both a morphosyntactic and pragmatic level. Although often used without further ado, the term *complexity* needs a definition to clarify what the following discussion is about. Investigating the linguistic expressions of locative and existential predication, I talk about formal complexity here, i.e., about linguistic units, and not functional complexity. Following Rescher (1998) and Karlsson et al. (2008), I conceive complexity as being measured by (1) the number of units included in an item (*hippopotamus* is phonologically more complex than *frog*; *the cute cat* is morphosyntactically more complex than *cat*) and (2) the variety of units included in an item (*was going* is more complex than *went*, because it expresses progressive aspect in addition to past tense). Functionally, a high degree of formal complexity leads to high salience in discourse (Boswijk & Coler 2020).

Following this approach to linguistic complexity, there is ample evidence that existential items and existential clauses, as a rule, are more complex than their locative counterparts. English and French provide good initial examples of this pattern. Whereas locative clauses contain forms of the copula verbs *be* and *être*, respectively, it is the analytic constructions *there is* and *il y a*, respectively, in existential predication. In either case, expletive elements make the existential construction more complex than the locative construction. As for the languages under consideration here, a similar observation can be made for those languages which include existential items in existential clauses, i.e. Chulym Turkic, Yeniseian and Northern Samoyedic.[15] In Chulym Turkic, locative clauses display a zero copula, whereas existential clauses include the existential item *par* (56–57). The Yeniseian languages generally function likewise; however, the person and number of the figure may be cross-referred in locative clauses, as in Ket (58–59). In either case, the existential clause is more complex since it contains more free morphemes and more phonetic material.

(56) Chulym (Turkic, Northeastern; Filchenko et al. 2012: 204–205)

ämdä	*olar*	*kat-tiir-i*	*äp-teer-in-dä.*
now	3PL	wife-PL-POSS3	house-PL-POSS3-LOC

‘Now, they [and] their wives are in their house.’

[15] I leave out the cases of existential takeover here because the degree of (morphological) complexity is certainly the same in either type of predication if there is the same linking item.

(57) Chulym (Turkic, Northeastern; Filchenko 2016-2019: KondiyakovAF_04July2016_Interview-1.75)

pis-tiŋ	*al-ïvs-ta*	*kömäs*	*koɣur*	*kizi-lär*	***par.***
1PL-GEN	village-POSS1PL-LOC	few	lazy	person-PL	**EX**

'In our village, there are few lazy people.'

(58) Central Ket (Yeniseian; Dul'zon 1971: 122)

ət	*qa-**reŋ**,*
1PL.PRO	at.home-**1PL**

'We are at home.'

(59) Southern Ket (Yeniseian; Siberian Lang: glosses_kel05_baldingm_mordushka_0-29)

is',	*is'*	*χat*	***us'en'.***
fish	fish	there	**EX**

'There is fish there.'

The Northern Samoyedic data are slightly more complex to analyse. In Nganasan, for example, locative clauses contain the copula verb *i-*, whereas existential clauses are formed with the existential verbs *təi-* and *tənij-* (Wagner-Nagy 2019: 354–355; 357). From a phonetic point of view, the existential verbs are clearly more complex than the copula verb. Additionally, the existential verb traces back to the combination of the demonstrative stem *tə-* and the demonstrative adverb *təni* 'there', respectively, with the copula verb *i-*, so existential clauses are actually equative clauses from a diachronic point of view (60). Equative clauses, in turn, are a typical means for expressing existential clauses, as, e.g. in Icelandic (isl; Indo-European, Germanic) (61), where they are opposed to locative clauses formed with the simple copula verb *vera* (Creissels 2019: 79–80). Summing up this argumentation, Nganasan – and, similarly, the other Northern Samoyedic languages – also provide evidence that existential clauses are more complex than locative clauses.

(60) Nganasan (Uralic, Samoyedic; INEL Nganasan Corpus: TKF_031118_War_nar.50)

tahar͡iabə	*təndə*	*sʲiti*	*bəŋgüʔtüə*	*təi-ču*	*(<tə-i-ču).*
now	there	two	burrow	EX-AOR.3SG	(that-be-AOR.3SG)

'Now, there are two burrows.' (< lit. 'Now, that is two burrows there.')

(61) Icelandic (Indo-European, Germanic; Creissels 2019: 79)
Það eru mys í baðkerinu.
that are mice in bathtub
'There are mice in the bathtub.' (lit. 'That are mice in the bathtub.')

The languages discussed so far distinguish locative from existential clauses by the linking element, which allows a comparably simple analysis of their complexity. But, as Creissels (2019) mentioned, there are many languages in which the linking element is one and the same in either construction. So, the linking element itself and its syntactic structure cannot indicate the complexity of the construction. In the analysed language sample, the Ob-Ugric (Khanty, Mansi), the Southern Samoyedic (Selkup, Kamas), two Turkic (Dolgan, Sakha) and the Tungusic (Evenki, Even) languages display this type, i.e. there is no morphosyntactic distinction of locative and existential predications. As a case in point, Northern Mansi displays the present-tense, third-person singular form of the copula verb in either sentence of (62–63).

(62) Northern Mansi (Uralic, Ob-Ugric; OUDB Northern Mansi Corpus: Text 1238, 016)
eːkʷa piːrisʲ jun ***oːl-i.***
Ekwa Piris at.home **be-PRS.3SG**
'Ekwa Piris is at home.'

(63) Northern Mansi (Uralic, Ob-Ugric; OUDB Northern Mansi Corpus: Text 1237, 003)
tit aːs waːta-t uːs ***oːl-i.***
here Ob bank-LOC town **be-PRS.3SG**
'There is a town on the bank of the Ob [river].'

Only word order distinguishes the two readings here, which evokes the question of whether there is evidence to analyse the word order in the existential clause (63) as more complex than the word order in the locative clause (62). When looking barely at the morphosyntax of these clauses, there is no indication that this would be the case. However, their information structure also points towards the existential clause being more complex than the locative clause. As discussed in Section 2.1, locative clauses usually show predicate or argument focus patterns, whereas existential clauses

exhibit sentence focus, which is why they are suitable for introducing new referents into the discourse. Following Lambrecht (1994: 222, 234–235) and Bentley et al. (2015: 43–44), sentence focus structures necessarily have the subject of the clause included in the focus domain. As a corollary, the subject is not topical. Given that subjects generally tend to be topical, yielding a parallel subject-predicate and topic-comment structure, it can be argued that sentence focus structures are more complex from an information-structural point of view than predicate or argument focus structures. Applied to the Mansi examples, this means that the topic-comment and subject-predicate structures are aligned in (62) (*e:k^{w}a pi:ris^{j}* 'Ekwa Piris' is both subject and topic), whereas in (63), *tit a:s wa:tat* 'on the bank of the Ob' is the topic, and *u:s* 'town' is the subject. Understanding information structure as a part of the syntax of the clause, (63) is, thus, syntactically more complex than (62).

Whereas the languages under investigation here provide only indirect evidence for this assumption, other languages are more expressive in this respect. One example of them is Finnish. In Finnish existential clauses, as well as in other clauses with sentence focus, a plural subject is marked with the partitive case and does not agree with the verb, as displayed in (64a) and (64c). In the correlating predicate focus structures, in turn, the subject stands in the nominative case and agrees with the verb, as displayed in (64b) and (64d). Consequently, the more complex information structure of (64a) and (64c) is also reflected in their morphosyntactic realisation, namely by additional case marking and missing person-number agreement.

(64) Finnish (Uralic, Finnic; personal knowledge)

a.	*Pöydä-llä*	*on*	*kirjo-j-a.*	EXISTENTIAL CLAUSE,
	table-ADE	be.3SG	book-PL-PTV	SENTENCE FOCUS
	'There are books on the table.'			
b.	*Kirja-t*	*o-vat*	*pöydä-llä.*	LOCATIVE CLAUSE,
	book-PL.NOM	be-3PL	table-ADE	PREDICATE FOCUS
	'The books are on the table.'			
c.	*Kadu-lla*	*leikki-i*	*laps-i-a.*	VERBAL INTRANSITIVE CLAUSE,
	street-ADE	play-3SG	child-PL-PTV	SENTENCE FOCUS
	'There are children playing in the street.'			
d.	*Lapse-t*	*leikki-vät*	*kadu-lla.*	VERBAL INTRANSITIVE CLAUSE,
	child-PL.NOM	play-3PL	street-ADE	PREDICATE FOCUS
	'The children are playing in the street.'			

When now combining the criteria “text frequency” and “generalisation in co-expression patterns” with the criterion “complexity” to assess the markedness of locative and existential predications, they contradict each other. The higher text frequency of existential predications predicts that they represent the unmarked item of a markedness opposition. In contrast, their higher complexity indicates that they represent the marked item. Since both locative and existential patterns can be generalised in the case of co-expression, this criterion is not finally expressive for determining the unmarked item of a markedness opposition.

As a possible alternative to this ‘dilemma’, I argue that the observations do not contradict each other per se. Starting from the assumption that locative and existential predications share their propositional content, a language has the “task” to disambiguate the possible readings – locative vs existential – by the grammatical means the language has. There is no default strategy for this disambiguation, as Creissels (2019) convincingly shows, but a wide variation can be observed. Still, the discourse-pragmatic functions of existential predications, (re-)introducing referents and structuring a discourse, make them more salient than locative predications, so the linguistic expressions of existential predications are often, though not necessarily, more complex. Acknowledging the higher salience of existential predications, it is not surprising that they are more frequent than locative predications. As discussed above, existential predications are needed for structuring a discourse, whereas locative predications cannot fulfil this function.

Given this, these characteristics and distinctions of locative and existential predications do not need a markedness-based opposition as an explanation when understanding *markedness* as the presence or absence of a phonetic or semantic feature (see above). Instead, both types of predication share their propositional content but differ in their perspective structure, so they are discriminated against each other in the given linguistic context. Formally, this is often – but not necessarily – instantiated by means of information structure. So, the semantically identical locative and existential predications merely serve different pragmatic domains, which is again an argument against assuming a markedness-based opposition.

5. Typological implications

The essence of the discussion in the preceding sections is that locative and existential predications do not exhibit a markedness opposition. Therefore, locative predications

must not be regarded as the unmarked structure from which a typology of locative and existential predication starts. In other words, there must not be any a priori restrictions, which items may occur in either type of predication; especially, the appearance of existential items in locative predication must be acknowledged. Still, they should be accounted for as two separate domains since their functional load heavily differs.

To capture all essential aspects of the linguistic expression of locative and existential predications, I propose a two-layered typology of locative and existential predication. At the first layer, the expressions of locative and existential predication are analysed and typologised independently. Here, the linking element is central to the typology since it displays most variation, whereas the coding of the figure and ground referents is already predetermined by the spatial figure-ground relationship expressed. Thus, Ket locative predications may be typologised as containing a zero copula with pn-agreement at the figure, whereas Ket existential predications may be analysed as containing an existential item/copula. Obviously, the typology itself needs a lot of elaboration, which may take the approaches discussed in Section 2.2 as its starting point. From a comparative and typological point of view, this layer of the typology makes a cross-linguistic study of the comparative concepts of "locative predication" and "existential predication" possible.

So far, the typology does not account for the tight interaction of locative and existential predication, formally reflected in many languages. Taking up the very initial step of Creissels' (2019) typology, the second layer of my proposed typology shall analyse how the linguistic expressions of locative and existential predications in a language are related to each other. In other words, it shall be analysed whether this language has co-expression patterns and, if applicable, how the morphosyntactic ambiguity is resolved, e.g. by word order permutations, different intonation contours or the like. Additionally, if needed for the relevant research purpose, it might be analysed whether the observed structure is existential in its origin, e.g. sharing its structure with generic existentials, having the same structure as other types of non-verbal predication, etc. To put this in a nutshell, the second layer of the typology shall make a cross-linguistic analysis of the (non-)co-expression patterns of locative and existential predication possible.

From my point of view, such a two-layered typology can overcome the methodological problems observed in Section 2.2. Most importantly, it is unbiased towards any linguistic expression of locative and existential predication. Furthermore,

it does not use the highly debated concept of markedness but relies on a functional approach to the semantics and pragmatics of the discussed predication types. Consequently, the typology – when appropriately elaborated – should be able to capture any language data showing instances of locative and existential predications, not excluding any of them by a priori restrictions.

6. Conclusions and further outlook

The initial observation of this paper was that fourteen Siberian languages exhibit existential items and structures in the linguistic expression of locative predication, that is, in locative clauses. This phenomenon was called "existential takeover". Subsequently, it was argued that a markedness-based approach to locative and existential predication is not appropriate since it makes contradictory predictions. The zero hypothesis of marked existential predications would predict the spread of the locative clause patterns in the case of formal neutralization, which is tackled by the instances of existential takeover discussed in this paper. Additionally, and perhaps even more importantly, it was shown that the parameters of textual frequency and complexity contradict each other, when being applied to determining the (un)markedness of locative and existential predications. Existential predications are, as a rule, significantly more frequent than locative predications which would entail them being the unmarked item of the opposition, whereas their higher complexity would entail them being the marked item. Consequently, it is hardly feasible to account for locative and existential predication in terms of a markedness opposition.

Instead, locative and existential predications share their propositional content, either of them expressing the presence or absence of a figure in a ground. Pragmatically, they are distinguished by opposing perspectivisation patterns, which result in a different information-structural configuration. Locative predications are perspectivised from the figure to the ground and exhibit predicate or argument focus, whereas existential predications provide a perspective on the whole situation, which correlates with sentence focus.

Following this, it is argued that a general typology of locative and existential predication must not assume either type as primary or unmarked. As described in Section 5, I propose a two-layered model of such a typology. The first layer describes the linguistic structures themselves and the second layer describes their interplay and possible co-expression patterns. Obviously, this proposed typology needs much

further elaboration. The approaches presented in Section 2.2 can serve well as a starting point but must definitely be fed with sufficient cross-linguistic data.

Finally, I would like to draw attention to a general issue in linguistic typology, namely the co-existence of two or more linguistic structures for the expression of a given comparative concept. Take, for example, the Chulym Turkic examples (24–27) in Section 3.3, which showed that negative locative and existential predications are formed by the negative existential *čok* ~ *čoɣul* in TAME-unmarked forms, but by the copula *pol-* ~ *bol-* in TAME-marked forms. From my point of view, both patterns must be included in a typology since neither context is a "better" representative of Chulym Turkic. Instead, one can posit a TAME-based split, as done by, e.g. Stassen (1997). In this context, the observed languages point to another feature, which cannot be analysed in detail here but should probably be acknowledged in a general typology of locative and existential predication. Almost all of them display a polarity split to some extent, so affirmative and negative structures are formed differently. As a case in point, the Southern Samoyedic languages Selkup and Kamas use a copula in affirmative locative and existential clauses but a negative existential item in their negative counterparts (see Section 3.3). To my knowledge, such a polarity split is not systematically acknowledged yet in the study of locative and existential predication. Still, it is probably a relevant factor judging on the base of the analysed Siberian languages. However, this goes far beyond the scope of this paper and must be postponed for further research.

In summary, the paper at hand may serve two independent but interwoven purposes. First, it adds knowledge to the description of locative and existential predication in Siberian languages. Second, it argues to clarify some theoretical issues of locative and existential predication and may, thus, serve as the starting point for the design and development of a general typology of locative and existential predication.

Acknowledgements

This publication was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 490822200. I would also like to thank the two anonymous reviewers for their very helpful comments, which led me to re-think some aspects and make the analysis more comprehensive. Additionally, I would like to thank Josefina Budzisch and Eva Schleitzer for fruitful

discussions and critical proofreading. It goes without saying that any remaining errors or uncertainties are my own.

Abbreviations

1 = 1st person
2 = 2nd person
3 = 3rd person
ABL = ablative
ACC = accusative
ADE = adessive
ADM = admonitive
AOR = aorist
AUX = auxiliary
CO = co-affix
COND = conditional
CVB = converb
DAT = dative
DIST = distal
DU = dual
EMPH = emphatic
EVID = evidential
EX = existential
FUT = future
GEN = genitive
HAB = habitual
INDEF = indefinite
INF = infinitive
LAT = lative
LOC = locative
MOM = momentaneous
MULT = multiple action
N = neuter
NEG = negative
NOM = nominative
OPT = optative
ORD = ordinal numeral
PL = plural
POSS = possessive
PROL = prolative
PRS = present
PST = past
PTCP = participle
PTV = partitive
R = Russian copy
RFL = reflexive
SEQ = sequential
SG = singular
VOC = vocative

References

Literature

Alekseev, Nikolaj A. 1995. *Predanija, legendy i mify sacha (jakutov)* [Traditions, legends and myths of the Sakha (Yakuts)] (Pamjatniki Fol'klora Narodov Sibiri i Dal'nego Vostoka 9). Novosibirsk: Nauka.

Ameka, Felix K. & Levinson, Stephen C. 2007. Introduction: The typology and semantics of locative predicates: posturals, positionals, and other beasts. *Linguistics* 45(5/6). 847–871. https://doi.org/10.1515/LING.2007.025.

Anderson, Gregory D.S. 2006. Towards a typology of the Siberian linguistic area. In Matras, Yaron & McMahon, April & Vincent, Nigel (eds.), *Linguistics areas. Convergence in historical and typological perspective*, 266–300. Basingstoke: Palgrave Macmillan.

Baranova, Vlada V. & Fedotov, Maksim L. & Oskolskaja, Sofia A. 2021. Expressing absence in the Turkic languages of the Volga-Kama sprachbund: Chuvash and Bashkir. *Tomsk Journal of Linguistics and Anthropology* 4(34). 9–31. https://doi.org/10.23951/2307-6119-2021-4-9-31.

Bentley, Delia & Ciconte, Francesco Maria & Cruschina, Silvio. 2015. *Existentials and locatives in Romance dialects of Italy*. Oxford: Oxford University Press.

Benzing, Johannes. 1955. *Lamutische Grammatik. Mit Bibliographie, Sprachproben und Glossar.* [Lamut grammar. With a bibliography, speech samples and a glossary]. (Akademie Der Wissenschaften Und Der Literatur. Veröffentlichungen Der Orientalischen Kommission 6). Wiesbaden: Steiner.

Böhtlingk, Otto. 1851. *Über die Sprache der Jakuten. Theil 1. Einleitung. Jakutische Texte. Jakutische Grammatik.* [About the language of the Yakuts. Part 1. Introduction. Yakut texts. Yakut Grammar.]. St. Petersburg: Buchdruckerei der Kaiserlichen Wissenschaften.

Borschev, Vladimir & Partee, Barbara H. 2002. The Russian genitive of negation: theme-rheme structure or perspective structure? *Journal of Slavic Linguistics* 10. 105–144.

Boswijk, Vincent & Coler, Matt. 2020. What is salience? *Open Linguistics* 6(1). https://doi.org/10.1515/opli-2020-0042.

Bybee, Joan L. 2011. Markedness: Iconicity, economy, and frequency. In Song, Jae Jung (ed.), *The Oxford handbook of linguistic typology*, 131–147. Oxford: Oxford University Press.

Clark, Eve. 1978. Locationals: Existential, locative and possessive constructions. In Clark, Eve & Greenberg, Joseph (eds.), *Universals of human language. Vol. 4. Syntax*, 85–126. Stanford [CA]: Stanford University Press.

Creissels, Denis. 2019. Inverse-locational predication in typological perspective. *Italian Journal of Linguistics* 31(2). 37–106. https://doi.org/10.26346/1120-2726-138.

Croft, William. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge: Cambridge University Press.

Däbritz, Chris Lasse. 2021. *Topik, Fokus und Informationsstatus: Modellierung am Material nordwestsibirischer Sprachen.* [Topic, focus and information status: Modelling on the basis of material from North-Western Siberian languages] (Language, Context and Cognition 17). Berlin, Boston: De Gruyter.

Däbritz, Chris Lasse. 2022. *A grammar of Dolgan. A Northern Siberian Turkic Language of the Taimyr Peninsula*. Leiden: Brill.

Däbritz, Chris Lasse & Wagner-Nagy, Beáta. 2024. Existential, locative and possessive predication in Kamas. *Journal of Uralic Linguistics* 3(1), 4-29. https://doi.org/10.1075/jul.00024.dab.

Däbritz, Chris Lasse. 2024. Existential, locative and possessive predication in Siberian Turkic languages: Co-expression, generalisation and spreading patterns. *Turkic Languages* 28. 70–110.

Dryer, Matthew S. 2007. Clause types. In Shopen, Timothy (ed.), *Language typology and syntactic description. 2nd edition. Volume I: Clause structure*, 224–275. Cambridge: Cambridge University Press.

Dul'zon, Andrej P. 1966. *Ketskie skazki.* [Ket folk tales]. Tomsk: Tomskij gosudarstvennyj pedagogičeskij institut.

Dul'zon, Andrej P. 1971. Materialy po ketskoj dialektologii. [Materials on Ket dialectology]. In Dul'zon, Andrej P. (ed.), *Jazyki i toponimija Sibiri.* [Languages and toponymy of Siberia], vol. 3, 119–158. Tomsk: Izdatel'stvo Tomskogo Universiteta.

Emel'janov, Nikolaj V. & Smirnov, Jurij I. 2008. *Jakutskie narodnye skazki* [Yakut folk tales] (Pamjatniki Fol'klora Narodov Sibiri i Dal'nego Vostoka 27). Novosibirsk: Nauka.

Eriksen, Pål Kristian. 2011. 'to not be' or not 'to be': The typology of negation of non-verbal predicates. *Studies in Language* 35(2). 275–310. https://doi.org/10.1075/sl.35.2.02eri.

Erteschik-Shir, Nomi. 2019. Stage topics and their architecture. In Molnár, Valéria & Egerland, Verner & Winkler, Susanne (eds.), *Architecture of topic* (Studies in Generative Grammar 136), 223–248. Berlin, Boston: De Gruyter Mouton.

Filchenko, Andrey & Potanina, Olga S. & Bajdak, Aleksandra V. & Fedotova, N. L. & Glazunov, Pavel Ju. & Gusev, Valentin Ju. & Kim, Antonina A. & Krjukova, Elena A. & Lemskaja, Valerija M. & Maksimova, N. P. & Tokmašev, Denis M. (eds.) 2010. *Annotated folklore and daily prose texts in the languages of the Ob-Yenisei area*. Vol. 1. Tomsk: Veter.

Filchenko, Andrey & Potanina, Olga S. & Bajdak, Aleksandra V. & Gusev, Valentin Ju. & Kim, Antonina A. & Krjukova, Elena A. & Lemskaja, Valerija M. & Maksimova, N. P. (eds.) 2012. *Annotated folklore and daily prose texts in the languages of the Ob-Yenissei linguistic area*. Vol. 2. Tomsk: Vajar.

Filchenko, Andrey & Potanina, Olga S. & Bajdak, Aleksandra V. & Fedotova, N. L. & Gusev, Valentin Ju. & Kim, Antonina A. & Kovylin, Sergej V. & Krjukova, Elena A. & Kurganskaja, Ju. V. & Lemskaja, Valerija M. & Maksimova, N. P. & Tokmašev, Denis M. & Tonojan, M. N. (eds.) 2013. *Annotated folklore and daily prose texts in the languages of the Ob-Yenissei linguistic area*. Vol. 3. Tomsk: Vajar.

Filchenko, Andrey & Potanina, Olga S. & Bajdak, Aleksandra V. & Brykina, Maria & Fellan, P. M. & Il'jašenko, I. A. & Kim, Antonina A. & Kovylin, Sergej V. & Krjukova, Elena A. & Kurganskaja, Ju. V. & Lemskaja, Valerija M. & Maksimova, N. P. & Tokmašev, Denis M. & Varda, Viktor E. & Wagner-Nagy, Beáta (eds.) 2015. *Annotated folklore and daily prose texts in the languages of the Ob-Yenissei linguistic area*. Vol. 4. Tomsk: Vajar.

Filchenko, Andrey & Potanina, Olga S. & Il'jašenko, I. A. & Kim, Antonina A. & Kovylin, Sergej V. & Krjukova, Elena A. & Lemskaja, Valerija M. & Maksimova, N. P. & Tokmašev, Denis M. & Wagner-Nagy, Beáta (eds.) 2017. *Annotated folklore and daily prose texts in the languages of the Ob-Yenissei linguistic area*. Vol. 5. Tomsk: Vajar.

Filchenko, Andrey & Brykina, Maria & Kovylin, Sergej V. & Krjukova, Elena A. & Lemskaja, Valerija M. & Maksimova, N. P. & Nefedov, A. V. & Tokmašev, Denis M. & Wagner-Nagy, Beáta (eds.) 2020. *Annotated folklore and daily prose texts in the languages of the Ob-Yenissei linguistic area*. Vol. 6. Tomsk: Agraf-Press & Vajar.

Filchenko, Andrey & Kim, Antonina A. & Kovylin, Sergej V. & Krjukova, Elena A. & Lemskaja, Valerija M. & Maksimova, N. P. & Nefedov, A. V. & Tokmašev, Denis M. (eds.) 2021. *Annotated folklore and daily prose texts in the languages of the Ob-Yenissei linguistic area*. Vol. 7. Tomsk: Agraf-Press & Vajar.

Francez, Itamar. 2007. *Existential propositions*. Stanford [CA]: Stanford University. (Doctoral Dissertation.)

Freeze, Ray. 1992. Existentials and other locatives. *Language* 68(3). 553–595.

Georg, Stefan. 2007. *A descriptive grammar of Ket (Yenisei-Ostyak). Part 1: Introduction, phonology, morphology*. Folkestone: Global Oriental.

Givón, Talmy. 1978. Negation in language: Pragmatics, function, ontology. In Cole, Peter (ed.), *Syntax and Semantics. Vol. 9. Pragmatics*, 69–112. New York: Academic Press.

Greenberg, Joseph. 2005[1966]. *Language universals. With special reference to feature hierarchies.* With a preface by M. Haspelmath (Janua Linguarum – Minor 59). Berlin, New York: Mouton de Gruyter.

Haspelmath, Martin. 2006. Against Markedness (And What to Replace It With). *Journal of Linguistics* 42(1). 25–70. https://doi.org/10.1017/S0022226705003683.

Haspelmath, Martin. 2022. *Nonverbal clause constructions*. Submitted manuscript. https://ling.auf.net/lingbuzz/006673 (Accessed November 17, 2002).

Helimski, Eugen. 1997. *Die matorische Sprache. Wörterverzeichnis, Grundzüge der Grammatik, Sprachgeschichte* (Studia Uralo-Altaica 41). Szeged: Szegedi Tudományegyetem.

Hengeveld, Kees. 1992. *Non-verbal predication: Theory, typology, diachrony* (Functional Grammar Series 15). Berlin, New York: Mouton de Gruyter.

Hölzl, Andreas. 2015. A typology of negation in Tungusic. *Studies in Language* 39(1). 117–157. https://doi.org/10.1075/sl.39.1.05hoe.

Janda, Laura A. 1996. Unpacking markedness. In Cased, Eugene H. (ed.), *Cognitive linguistics in the redwoods* (Cognitive Linguistics Research 6), 207–233. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110811421.207.

Janhunen, Juha. 1977. *Samojedischer Wortschatz. Gemeinsamojedische Etymologien* [Samoyedic lexicon. Common Samoyedic etymologies]. (Castrenianumin Toimitteita 17). Helsinki: Suomalais-Ugrilainen Seura.

Janhunen, Juha. 2003. Proto-Mongolic. In Janhunen, Juha (ed.), *The Mongolic languages*, 1–29. London: Routlegde.

Johanson, Lars. 2021. *Turkic.* Cambridge: Cambridge University Press.

Junghanns, Uwe. 2002. Informationsstrukturierung in slavischen Sprachen: Zur Rekonstruktion in einem syntax-zentrierten Modell der Grammatik [Information structuring in Slavic languages: On reconstruction within a syntax-centered grammar model]. Leipzig: University of Leipzig. (Habilitation Dissertation.)

Kahn, Charles H. 1966. The Greek verb ‘to be’ and the concept of being. *Foundations of Language* 2(3). 245–265.

Karakoç, Birsel. 2009. Notes on subject markers and copular forms in Turkish and in some Turkic varieties of Iran: A comparative study. *Turkic Languages* 13. 208–224.

Karlsson, Fred & Miestamo, Matti & Sinnemäki, Kaius. 2008. Introduction. The problem of language complexity. In Miestamo, Matti & Sinnemäki, Kaius & Karlsson, Fred (eds.), *Language complexity. Typology, contact, change* (Studies in Language Companion Series 94), vii–xiv. Amsterdam, Philadelphia: John Benjamins.

Koch, Peter. 2012. Location, existence, and possession: A constructional-typological exploration. *Linguistics* 50(3). 533–603. https://doi.org/10.1515/ling-2012-0018.

Kotorova, Elizaveta G. & Porotova, Tel’mina I. 2001. *Ketskie fol’klornye i bytovye teksty* [Ket folklore and everyday texts]. Tomsk: Tomskij gosudarstvennyj pedagogičeskij universitet.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus, and the mental representations of discourse referents* (Cambridge Studies in Linguistics 71). Cambridge: Cambridge University Press.

Lyons, John. 1967. A note on possessive, existential and locative sentences. *Foundations of Language* 3(4). 390–396.

McNally, Louise. 2011. Existential sentences. In Maienborn, Claudia & von Heusinger, Klaus & Portner, Paul (eds.), *Semantics: An International Handbook of Natural Language Meaning* (Handbücher Zur Sprach- Und Kommunikationswissenschaft 33), vol. 2, 1829–1848. Berlin: Mouton de Gruyter.

McNally, Louise. 2016. Existential sentences crosslinguistically: Variations in form and meaning. *Annual Review of Linguistics* 2. 211–231. https://doi.org/10.1146/annurev-linguistics-011415-040837.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective* (Empirical Approaches to Language Typology 31). Berlin: Mouton de Gruyter.

Milsark, Gary. 1974. *Existential Sentences in English*. Cambridge [Mass.]: Massachusetts Institute of Technology. (Doctoral Dissertation.)

Miščenko, Darja F. 2017. Neglagol'nye predloženija baškirskogo jazyka i sposoby vyraženija otricanija v nix [Non-verbal sentences of the Bashkir language and strategies for the expression of negation in them]. *Acta Linguistica Petropolitana* 13(1). 110–146.

Molnár, Valeria. 1991. *Das TOPIK im Deutschen und Ungarischen* [The TOPIC in German and Hungarian] (Lunder Germanistische Forschungen 58). Stockholm: Almqvist & Wiksell International.

Munkácsi, Bernát. 1892. *Vogul népköltési gyűjtemény* [Collection of Vogul folk poetry]. Vol. 1. Budapest: Magyar Tudományos Akadémia.

Munkácsi, Bernát. 1893. *Vogul népköltési gyűjtemény* [Collection of Vogul folk poetry]. Vol. 3. Budapest: Magyar Tudományos Akadémia.

Paducheva, Elena V. 2008. Locative and existential meaning of Russian быть. *Russian Linguistics* 32(3). 147–158.

Panova, Anastasia & Liljegren, Henrik. Forthcoming. Locative and existential predication contrasts in Gawarbati (Indo-Aryan) and the surrounding region. In Däbritz, Chris Lasse & Budzisch, Josefina & Basile, Rodolfo (eds.), *Locative and Existential predication: On forms, functions and neighbouring domains*. Berlin: Language Science Press.

Rescher, Nicholas. 1998. *Complexity. A philosophical overview*. New Brunswick, London: Transaction Publishers.

Sasse, Hans-Jürgen. 1987. The thetic/categorical distinction revisited. *Linguistics* 25(3). 511–580. https://doi.org/10.1515/ling.1987.25.3.511.

Sotavalta, Arvo. 1978. *Westlamutische Materialien* [Western Lamut materials] (Suomalais-Ugrilaisen Seuran Toimituksia 168). Edited by Harry Halén. Helsinki: Suomalais-Ugrilainen Seura.

Stassen, Leon. 1997. *Intransitive Predication* (Oxford Studies in Typology and Linguistic Theory). Oxford: Clarendon Press.

Steinitz, Wolfgang. 1967. *Dialektologisches und etymologisches Wörterbuch der ostjakischen Sprache*. Vol. 2. Berlin: Akademie-Verlag.

Steinitz, Wolfgang. 1975. Ostjakologische Arbeiten [Ostyakological works]. Vol. 1. Ostjakische Volksdichtung und Erzählungen aus zwei Dialekten. Texte [Ostyak Folk Poetry and Stories from two dialects. Texts]. Edited by Gert Sauer. Den Haag: Mouton.

Steinitz, Wolfgang. 1989. Ostjakologische Arbeiten [Ostyakological works]. Vol. 3. Texte aus dem Nachlass [Texts from the estate]. Edited by Gert Sauer, Renate Steinitz, Lieselotte Hartung, Petra Hauel, Brigitte Schulze. Berlin, New York: Mouton de Gruyter.

Talmy, Leonard. 1983. How language structures space. In Pick Jr., Herbert L. & Acredolo, Linda P. (eds.), *Spatial orientation. Theory, research, and application*, 225–282. New York, London: Plenum Press.

Tottie, Gunnel. 1991. *Negation in English Speech and Writing: A Study in Variation* (Quantitative Analyses of Linguistic Structure 4). San Diego [CA]: Academic Press.

Ubrjatova, Elizaveta I. & Korkina, Evdokija I. & Charitonov, Luka N. & Petrov, Nikolaj E. 1982. *Grammatika sovremennogo jakutskogo literaturnogo jazyka* [Grammar of the Modern Literary Yakut language]. Moskva: Nauka.

Vajda, Edward. 2009. The languages of Siberia. *Language and Linguistics Compass* 3(1). 424–440.

Veselinova, Ljuba. 2013. Negative existentials: A cross-linguistic study. *Rivista di Linguistica* 25(1). 107–145.

Veselinova, Ljuba. 2015. Special negators in the Uralic languages. In Miestamo, Matti & Tamm, Anne & Wagner-Nagy, Beáta (eds.), *Negation in Uralic languages* (Typological Studies in Language 108), 547–599. Amsterdam, Philadelphia: John Benjamins.

Veselinova, Ljuba & Hamari, Arja. 2022. Introducing the negative existential cycle. In Veselinova, Ljuba & Hamari, Arja (eds.), *The negative existential cycle* (Research on Comparative Grammar 2). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6306474.

Wagner-Nagy, Beáta. 2019. *A Grammar of Nganasan*. Leiden: Brill.

Waugh, Linda R. & Lafford, Barbara A. 2000. Markedness. In Booij, Geert E. & Lehmann, Christian & Mugdan, Joachim (eds.), *Morphologie* (Handbücher zur Sprach- und Kommunikationswissenschaft 17/1), 272–281. Berlin,New York: De Gruyter Mouton.

Werner, Heinrich. 1997. *Das Jugische (Sym-Ketische)* [Yugh (Sym-Ket)] (Veröffentlichungen Der Societas Uralo-Altaica 50). Wiesbaden: Harrassowitz.

Corpora and databases

DOBES collection "Even"

Natalia Aralova, Brigitte Pakendorf, Alexandra Lavrillier, Dejan Matic, Natalia Mikhailovna Golikova, Katharina Gernet, Tat'jana Vasil'evna Zakharova, Khristina Mikhailovna Zakharova, Matrena Gavrilovna Golikova (Pogodaeva), Viktoria Akhtamovna Lebedeva, Maria Petrovna Djachkovskaja, Mikhail Alekseevich Turantaev, Raisa Petrovna Kuzmina, Stepan Mikhailovič Lebedev, Arsen Timofeevich Slepcov, Dar'ja Mikhailovna Osenina, Evdokia Vasil'evna Semenovna, Luise Zippel, Maria Petrovna Lomovceva, NA, Natalia Ionovna Grigoreva, and Tat'jana Afanas'evna Zabolotskaja. 2007 - 2023. Collection "Even". The Language Archive. https://hdl.handle.net/1839/9cbb2743-a47f-4767-b5fe-3b7764854fb3 (Accessed 2023-07-03).

ELAR Melets Chulym collection

Filchenko, Andrey. 2016–2019. *Comprehensive documentation and archiving of Teleut, Eushta-Chat, and Melets Chulym: three areally adjacent critically endangered Turkic languages of Siberia*. Endangered Languages Archive. http://hdl.handle.net/2196/00-0000-0000-0010-8981-B.

INEL Dolgan Corpus

Däbritz, Chris Lasse & Kudryakova, Nina S. & Stapert, Eugenie. 2022. *INEL Dolgan Corpus*. Version 2.0. https://doi.org/10.25592/uhhfdm.11165.

INEL Enets Corpus

Khanina, Olesya & Shluinsky, Andrey. In preparation. *INEL Enets Corpus*. (Unpublished.)

INEL Evenki Corpus

Däbritz, Chris Lasse & Gusev, Valentin. 2021. *INEL Evenki Corpus*. Version 1.0. https://hdl.handle.net/11022/0000-0007-F43C-3.

INEL Kamas Corpus

Gusev, Valentin & Klooster, Tiina & Wagner-Nagy, Beáta. 2019. *INEL Kamas Corpus*. Version 1.0. http://hdl.handle.net/11022/0000-0007-DA6E-9.

INEL Nenets Corpus

Budzisch, Josefina & Wagner-Nagy, Beáta. In Preparation. *INEL Nenets Corpus*. Version 1.0.

INEL Selkup Corpus

Brykina, Maria & Orlova, Svetlana & Wagner-Nagy, Beáta. 2021. *INEL Selkup Corpus*. Version 2.0. https://hdl.handle.net/11022/0000-0007-F4D9-1.

INEL Nganasan Corpus

Brykina, Maria & Gusev, Valentin & Szeverényi, Sándor & Wagner-Nagy, Beáta. In preparation. *INEL Nganasan Corpus*. (Unpublished. Predecessor version published under http://hdl.handle.net/11022/0000-0007-C6F2-8).

Ob-Ugric Database

Skribnik, Elena & Riese, Timothy (eds.). 2014. *Ob-Ugric database: analysed text corpora and dictionaries for less described Ob-Ugric dialects*. www.oudb.gwi.uni-muenchen.de (Accessed July 5, 2023).

Siberian Lang

Kazakevich, Olga A. (ed.). 2012. *Minority languages of Siberia as our cultural heritage*. http://siberian-lang.srcc.msu.ru/en (Accessed July 5, 2023).

Yakut-Russian Code-Switching Corpus

Petukhova, Anna A. & Sokur, Elena O. 2021. *Yakut-Russian Corpus of Code-Switching*. Moscow: International Laboratory of Language Convergence, Higher School of Economics. http://lingconlab.ru/cs_yakut (Accessed July 5, 2023).

CONTACT

chris.lasse.daebritz@uni-hamburg.de; daebritz@wissenschaftsrat.de

Linguistic Typology at the Crossroads 4-2 (2024): 125-157 ISSN 2785-0943

Relativization strategies and sociolinguistic variation in spoken Italian: a typological account

SILVIA BALLARÈ

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

Submitted: 28/10/2024 Revised version: 18/12/2024
Accepted: 20/12/2024 Published: 23/01/2025

Abstract

In this paper I aim at describing and analysing relative clauses in a corpus of spoken Italian. In the first section, I provide an overview of the relativization strategies in Italian, also taking into account non-standard varieties; then, I briefly discuss the sociolinguistic characterization of the sub-standard area of contemporary Italian. In § 2, I introduce the selected corpus and its characteristics, also explaining the methodologies adopted for data extraction and annotation. Then, in § 3, the results of the analysis are presented. The distributions of the different strategies and the outputs of a statistical analysis show the different importance assumed by both linguistic and extralinguistic factors and enable the explanation of the observed variability. Finally, in § 4, some general conclusions are drawn.

Keywords: relative clauses; sociolinguistic variation; language variation; spoken Italian.

1. Framework

Relative clauses are a widely studied topic in linguistics; also recently, much attention has been devoted to these structures from different perspectives (see e.g. Alexiadou et al. 2000; Kidd 2011; Henderey 2012; Ackerman & Nikolaeva 2013; Cinque 2020). Even in Italian, the topic has been discussed at length in the literature (see below). Relative clauses in Italian can be realized through an array of different strategies:

https://doi.org/10.6092/issn.2785-0943/20572

speakers have multiple options and both simplification and complexification processes come into play.

In this contribution, I discuss the behavior of relativization strategies of all grammatical relations in a small corpus of informal spoken Italian, involving speakers with different social characterizations. Linguistic and extralinguistic factors will be taken into account to discuss and explain the behavior of these structures in spoken data. In addition, the analysis will be conducted by adopting classificatory categories and notions typical of linguistic typology, given that "the patterns of variation and change found in [...] a particular language are in many cases simply instances of patterns of variation and change found across languages" (Croft 2022: 27).

The sociolinguistic analysis is intertwined with the adoption of typological theoretical tools to build a bridge between intralinguistic and interlinguistic variation (see Inglese & Ballarè 2023 inter al.). The analysis of structural differences displayed by varieties of the same language in a typological perspective on the one hand shows that non-standard variants are not to be considered mere "accidents" as are well attested in other languages and, on the other, it allows for crosslinguistic comparisons.

In the first part of this section, I present relativization strategies in Italian; in the second, adopting a sociolinguistic perspective, I introduce the Italian sub-standard area.

1.1. Relative clauses in Italian

In standard Italian, relative clauses can be realized through different strategies (see Serianni 2010 [1989]: 217-240). Nominatives and accusatives in non-restrictive relative clauses can be introduced by ART. + *quale* ('which'), which must be inflected to display gender/number agreement with the antecedent, as in (1). The same grammatical relations (both in restrictive and non-restrictive relative clauses) can be expressed with the invariable *che*[1] ('that'), as in (2). The first strategy is more formal,

[1] Some authors (see Cristofaro & Giacalone Ramat 2007) consider *che* ('that') as part of a morphological paradigm composed of two cells filled with *che* ('that') and *cui* ('which'), due to diachronic reasons. However, *che* ('that'), unlike *cui* ('which') that behave almost exclusively as a relativizing element in the standard variety, can be used with different functions. In fact, it can be employed, for example, to introduce completive clauses and adverbial subordinates, and, in the literature, it is often considered to be a "general subordinator" (also 'multifunctional *che*', see below). Furthermore, the two elements are placed in two different stages in the "pronominality cline" (lit. *cline di pronominalità*) proposed by Fiorentino (1999: 164). Because of the high polyfunctionality of *che* ('that') and, consequently,

and it is typically attested in highly controlled productions, while the latter is more neutral from a sociolinguistic point of view, as it occurs both in high and low productions.

(1) *Marco parla a Giulia, la quale dorme*
Marco talk:PRS.3SG to Giulia(SG.F) DEF:SG.F REL sleep:PRS.3SG
'Marco talks to Giulia, who sleeps.'

(2) *Marco parla a Giulia, che dorme*
Marco talk:PRS.3SG to Giulia REL sleep:PRS.3SG
'Marco talks to Giulia, who sleeps.'

All the other grammatical relations can be realized by the means of a preposition that expresses the function of the antecedent in the subordinate clause followed by the invariable *cui* ('which')[2], as in (3), or by the inflected form of ART. + *quale* ('which'), as in (4). *Cui,* even if not preceded by any preposition, can be used to express genitive if placed inside the noun phrase between the article and the noun; however, this use is quite rare and attested only in highly formal productions.

In addition, Italian, as many other European languages (see Murelli 2011: 184), has a dedicated form to relativize locative values, i.e. *dove* ('where'), as shown in (5). Some spatial values, such as the ablative, can be expressed by combining a preposition with *dove*, as in *da dove* 'from where'. Lastly, most grammars consider standard the employ of *che* ('that') to relativize a temporal value, as in (6).

(3) *La ragione per cui sono in ritardo è*
DEF:SG.F reason:SG.F for REL be:PRS.1SG late be:PRS.3SG
il maltempo
DEF:SG.M bad.weather:SG.M
'The reason why I am late is bad weather.'

differences in the breadth of functional domains between *che* ('that') and *cui* ('which'), these two elements are treated as independent. By favoring a synchronic approach, in fact the (historical) opposition of these two elements is in the process of being lost or, at the very least, weakened.

[2] Please note that, in this paper, *cui* and *quale* are both translated as *which*. However, as illustrated in this section, if preceded by a preposition, they can be both be employed also as *whom*, while *cui* can be used also as *whose*. Due to differences in the relativizing strategies between Italian and English, the translation can be misleading.

(4) *La* *ragazza* *della* *quale* *ti* *ho parlato*
DEF:SG.F girl:SG.F of.DEF:SG.F REL DAT.2SG speak:PST.1SG
è *Anna*
be:PRS.3SG Anna
'The girl I spoke to you about is Anna.'

(5) *La* *città* *dove* *vivo* *è* *Milano*
DEF:SG.F city:SG.F REL live:PRS.1SG be:PRS.3SG Milan
'The city where I live is Milan.'

(6) *Il* *giorno* *che* *ti* *ho* *conosciuto* *pioveva*
DEF:SG.M day:SG.M REL DAT.2SG AUX.1SG know:PST.PTCP rain.PST.3SG
'The day I met you it was raining.'

The whole paradigm of relativization strategies in standard Italian is summarized in table (1).

	no agreement with head noun	**agreement with head noun**
NOM. and ACC.	*che*	ART. + *quale*
LOC.	PREP. + *cui* (PREP. +) *dove*	PREP. + ART. + *quale*
TEMP.	PREP. + *cui* *che*	PREP. + ART. + *quale*
All other relations	PREP. + *cui*	PREP. + ART. *quale*

Table 1: Relativization strategies in standard Italian.

The relativization strategies attested in Italian can be categorized through the taxonomy proposed by Comrie & Kuteva (2013a, 2013b) in the World Atlas of Language Structures. In this perspective, standard Italian displays:

- two[3] *relative pronoun strategies* that involve ART. + *quale* and *cui*, in which the element is case marked by a preposition (or by its absence, since Italian does not have a dedicated adposition to express nominatives and accusatives) to

[3] One could also add the case of PREP. + *dove* ('where'), given that the preposition expresses the locative value.

indicate the role of the antecedent within the subordinate clause; more precisely, ART. + *quale* can be used to relativize nominative and accusatives, while PREPOSITION + *cui* ('which') or ART. + *quale* ('which') can be used to relativize all other grammatical relations.

- two *gap strategies*, realized through *che* ('that') and *dove* ('where'), where there is no overt case-marked reference to the head noun within the subordinate clause. *Che* ('that') is semantically empty (and, in fact, it is used to introduce different kinds of subordinate clauses such as the completive ones), while the meaning of *dove* ('where') is linked with locative values, just as English *where*; however, from a structural point of view, they are invariable and neither of them is case marked (by case or by an adposition).

The array of relativization strategies is much wider when taking into account non-standard varieties of Italian. First, some of the aforementioned elements have broadened their functional domain, and, thus, are used to express more values than in the standard variety. The invariable *che* ('that') is not rarely employed to relativize obliques, as in (7); *dove* ('where') is used with non-spatial antecedent and, sporadically and especially in interactions that involve speakers with low educational achievements, it can be used to relativize nominatives, as in (8) (for a detailed discussion see Ballarè & Inglese 2022).

(7) Alfonzetti 2022: 59 cit. in Cerruti (2017: 65)

Non	*c'*	*è*	*nessuno*	*che*	*posso*	*chiedere?*
NEG	there	be:PRS.3SG	nobody	REL	can:PRS.1SG	ask:INF

'Is there anyone I can ask?'

(8) Bernini 1989: 91

Nel	*greco*	*c'*	*è*	*un*	*dativo*	*dove*
in. DEF:M.SG	Greek.M.SG	there	be:PRS.3SG	INDEF:SG.M	dative:M	REL
può	*presentare*	*una*	*enne*	*finale*		
can.PRS.3SG	show:INF	INDEF:SG.F	n:SG.F	final:SG.F		

'In Greek there is a dative that can show a final *n*.'

Furthermore, as discussed in detail by Cerruti (2017), in non-standard varieties a wider range of structural possibilities is attested. It is worth noting at least 2 additional constructions.

The first one involves an invariable element - typically *che* ('that'), as in (9), but sporadically also *dove* ('where'), as in (10) - followed by a clitic pronoun which provides

information about the grammatical relation of the relativized element and in some cases agrees with it in terms of gender and number; this is true for datives, that show gender/number agreement, but not for locatives. When a nominative is relativized, given that Italian does not have subject clitics, a tonic pronoun is retained, as in (11). These cases are classified as *resumptive pronoun* by Comrie & Kuteva (2013a, 2013b).

(9) Alfonzetti 2002: 59 cit. in Cerruti (2017: 66)

I	*due*	*americani*	*che*	*gli*	*ho*	*aperto*
the:PL.M	two	american:PL.M	REL	DAT.3PL	AUX.1SG	open:PST.PTCP
l'	*ombrellone*					
DEF:SG.M	beach.umbrella:SG.M					

'The two Americans for whom I opened the beach umbrella.'

(10) KIParla Corpus, PTD012

Una	*strada […]*	*dove*	*ci*	*passa*	*molta*	*più*
INDEF:SG.F	street:SG.F	REL	LOC	pass.by:PRS.3SG	much:SG.F	more
gente						
people:SG.F						

'A street where much more people pass by.'

(11) Berretta 1993: 232 cit. in Cerruti (2017: 66)

c'	*era […]*	*Cesarini,*	*che*	*lui*	*all'*	*ultimo*
there	be:PST.3SG	Cesarini	REL	SUBJ.3SG	at.DEF:SG.M	last:SG.M
minuto	*faceva*	*sempre*	*goal*			
minute:SG.M	do:PST.3SG	always	goal			

'There was Cesarini, who always scored a goal at the last minute.'

Lastly, there are cases in which there is a *double encoding* (Murelli 2011) of the grammatical relation of the relativized element. More specifically, the construction consists of one inflected element (i.e. PREP. + ART. + *quale* or *cui*) followed by a clitic pronoun that re-expresses the grammatical relation of the antecedent in the subordinate clause, as in (12).

(12) itTenTen20 corpus

Sembravo	*un*	*bambino*	*a*	*cui*	*gli*	*era*
seem:PST.1SG	INDEF:SG.M	child:SG.M	to	REL	DAT.3SG	AUX.3SG

stato	*fatto*	*il*	*regalo*	*che*	*da*	*sempre*
PST.PTCP	do:PST.PTCP	DEF:SG.M	gift:SG.M	REL	from	always
aveva	*desiderato*					
AUX.3SG	desire:PST.PTCP					

'I looked like a child who had been given a gift that he had always desired.'

1.2. Sociolinguistic variation: the sub-standard area

As it has been shown, Italian displays a complex set of relativization strategies; in fact, different grammatical relations must be realized through different strategies and more strategies can be employed for the same grammatical relation.

Not surprisingly, these strategies display a different sociolinguistic characterization, and this setting lends itself well to numerous studies that, over the years, shed light over its variability and sociolinguistic variation (see Alisova 1965; Alfonzetti 2002; Fiorentino 1999; Cerruti 2016, 2017 inter al.). Suffice to say that in the seminal work authored by Berruto (2012) on sociolinguistic variation in contemporary Italian, relative clauses are emblematically selected as case study to give account for morphosyntactic variation (Berruto 2012: 48ff.). More specifically, the scholar creates a continuum in which he displays all the strategies by crossing two ordered dimensions: the first one is composed of different varieties of Italian- from the higher pole of written standard to the lower of *Italiano popolare* (lit. 'popular Italian', see below)- and the other one gives account of structural characteristics -from the synthetic to the analytic pole.

Regrettably, the persistent lack of freely accessible spoken corpora providing speakers' metadata that has characterized the Italian scenario over the years has led to difficulties in confronting systematically different varieties of Italian and in allowing for further reading of the data, especially from a quantitative perspective. This situation has been changing over the last few years, also thanks to the publication of the KIParla corpus (Mauri et al. 2019) which is a freely accessible resource consisting in spoken data accompanied by a large set of metadata (see § 2), allowing for the analysis of sociolinguistic variation.

In this study, I aim at describing and analyzing how relative clauses are realized in sub-standard productions. More specifically, informal spoken interactions involving speakers with different social characterizations will be taken into account. As it is well known, the informal style is the one in which speakers more easily distance themselves from the standard and, thus, allow us to investigate more in depth the behavior of deviant

strategies; furthermore, the social dimension will be considered because, traditionally, it has been considered highly explicatory to give account for sociolinguistic variation.

The social dimension has been of great relevance in identifying a very important variety within the architecture of contemporary, i.e. the so-called *Italiano popolare* (lit. 'popular Italian'). This variety has been identified in the Seventies (De Mauro 1970; Cortelazzo 1972) and it has been associated with speakers with low educational level that have an Italo-romance dialect[4] as a mother-tongue and employ Italian only in more controlled contexts, where the use of dialect would be strongly stigmatized. The visibility of *Italiano popolare* has greatly diminished in recent decades, and in the literature the scope of the label has been downplayed or the very existence of the variety has been denied (see Lepschy 2002; Renzi 2000, 2012). In support of these, following Berruto (2014: 278-279), two main arguments can be identified. One argues that there are no longer prototypical speakers of *Italiano popolare* and the other that the linguistic features that characterized *Italiano popolare* are to be considered generically sub-standard since they systematically appear in informal productions, regardless of the social characterization of the speakers.

The other main sub-standard variety is the so-called *colloquial Italian*, which is used in everyday, spoken but also written, interactions by speakers of various social characterization, included the ones with higher educational achievements (Berruto 2012 [1987]: 163; Ballarè 2024). From a sociolinguistic perspective, colloquial Italian is maximally relevant because it constitutes, along with *Italiano popolare*, the privileged place where linguistic innovations arise and thus the space in which ongoing variation can be observed.

In this paper, thanks to the analysis of relative constructions, it will be discussed if (and how) speakers with diverse social characterizations behave in different ways in informal spoken productions; this will allow us to discuss if, at least for relativization strategies, the sub-standard part of the architecture of contemporary Italian is homogenous or if there are relevant differences that allow us to distinguish different linguistic behaviors.

[4] Italo-romance dialects are different languages from Italian and *not* (geographical) varieties of Italian, in that they all derive from Latin and, thus, they display a structural distance from standard Italian that is similar to the one that can be found, for example, between Spanish and French. For a discussion regarding the structural characteristics of Romance languages, including Italo-romance dialects (such as the dialects of northern Italy -Benincà et al. 2016- and the one of southern Italy - Ledgeway 2016) see Ledgeway & Maiden (2016).

2. Data and methods

In this section, the ParlaTO corpus is briefly presented together with the choices that were made to identify two subcorpora; then, the methodology adopted in order to extract and code the data is explained.

2.1. ParlaTO corpus

The ParlaTO corpus (Cerruti & Ballarè 2021) is a module of the larger KIParla corpus (Mauri et al. 2019). It consists of semi-structured interviews collected in the urban area of Turin with speakers balanced by age group (16-29, 30-59, over 60) diversified by social characteristics (gender, educational achievement, occupation). The corpus consists of 48:51 hours of total recordings, 65 interviews and 552.461 tokens. For the purpose of this study, it is important to specify that the interviews, in the vast majority of cases, were conducted by students/researchers who were familiar with the informants (there are, for example, interviews involving relatives and friends) or otherwise in the presence of an intermediary (i.e., a person who knew both the interviewer and the interviewee, and that, by participating in the interaction, cooperated in making the exchange less controlled). In these interactions, speakers were asked for opinions about the city of Turin (about their neighborhood of residence, the change that had occurred over the years, etc.): the topic was selected because it was hypothesized that it might be of interest to the speakers and might engage them in expressing views and opinions. In addition, the exchanges almost always took place in locations selected by the interviewees themselves so that they could be more comfortable. Although the semi-structured interview is a rather codified type of interaction, due to the methodological choices made during the collection phase (see, e.g., Labov 1984: 32-33), overall, these can be considered as rather informal interactions.

In order to observe social variation, two subcorpora were created:

- Subcorpus L (166.540 tokens): semi-structured interviews with speakers with at the most a secondary school license; all available interviews within the corpus were taken, for a total of 12 interviews with 15 informants.
- Subcorpus H (169.376 tokens): semi-structured interviews with speakers with at least a high school diploma; in order to maximize the distance with the social characterization of the speakers of the other subcorpus, informants with a technical/professional school diploma were excluded. Through a randomization

of the selected interviews, a sample size similar to the previous one in terms of tokens was created, for a total of 18 interviews with 22 informants.

The parameter "educational achievement" was selected to divide the speakers into two groups and, consequently, create the two subcorpora exemplifying the social varieties under scrutiny. More specifically, in subcorpus L there are 9 speakers with a primary school license and 6 with a secondary school license; in subcorpus H, on the other hand, there are 7 speakers with a high school diploma, 6 college students and 9 college graduates. Among the available metadata, educational achievement was selected to create socially differentiated groups, as traditionally done in the literature (cfr. Berretta 1988). In fact, a different degree of education often correlates with morphosyntactic variation (see Berruto 1983 inter al.). Furthermore, note that *Italiano popolare* is identified *per* definitionem taking into account speakers' educational achievement.

Speakers of the two subcorpora, moreover, are also diversified by age group (and thus employment) and geographical origin, as shown in table (2).

	Subcorups L	**Subcorpus H**
	Age range	
21-30	1	6
31-40	0	6
41-50	0	2
51-60	0	3
61-70	4	3
71-80	5	1
Over80	5	1
	Occupation	
Retailers	0	2
Managers and directors	0	1
Laborers	1	0
Pensioners	14	4
University students	0	6
	Geographic origin	
North	8	19
Center	0	1
South and islands	7	2

Table 2: The social characterization of the speakers.

Looking at the values shown in the table, it is clear that in subcorpus L there are almost exclusively speakers over 60 years old (and, therefore, pensioners), half of whom were born in northern regions and the other half in southern regions. The picture is quite different in subcorpus H, where there are speakers of different age groups (from 21-30 to over80), who have various jobs and who in the vast majority of cases were born in northern regions. These differences must be linked to the fact that young people are, generally, higher educated and that a massive immigration from southern to northern regions took place in Italy from the 1950s to the 1970s.

The whole corpus has a small dimension and consists of 335.916 tokens; this is due to the fact that there were only 12 interviews with speakers with low educational achievements and, thus, already mentioned, in order to create a balanced sample, I decided to take into account a comparable number of tokens also for the subcorpus H. Furthermore, I was forced to use a rather small amount of data because the analysis of the scrutinized linguistic features required a very laborious and time-consuming manual work of data cleaning, given that the KIParla corpus is not tagged. For instance, to analyze the relative clauses realized through *che* ('that'), it was needed to manually select them among all the 6.072 occurrences of the aforementioned linguistic items in the corpus.

2.2. Data extraction and annotation

In order to detect all the relative clauses and given that the KIParla corpus is not morpho-syntactically annotated, all the occurrences of (PREP. +) ART. + *quale/i*, *cui*, *che* and *dove* were extracted; this led to a datafile composed of 6.973 occurrences that has been manually cleaned, ruling out:

a) Cases in which *che* ('that') and *dove* ('dove') were not used as relativizing elements but, for example, as complementizer for completive clauses or as interrogative pronoun/adverb in questions.

This selection was not always straightforward because of cases of the so called *che polivalente* (lit. 'multifunctional *che*', see Fiorentino 2011), that can introduce relative clauses or other subordinates. In order to disambiguate, all the cases in which *che* ('that'), according to the standard rules, could be replaced by another relativizing element were taken into account, as is the case of (13), in which, for example, *in cui* ('in which') could be used in place of *che* ('that').

(13) KIParla, PTD009

Non	*è*	*che*	*viviamo*	*in*	*Olanda,*	*che*	*con*	*quattro*
NEG	be:PRS.3SG	COMP	live:PRS.1PL	in	Holland	REL	with	four
gradi	*sotto*	*zero*	*prendi*	*la*	*bicicletta*			
degrees	below	zero	take:PRS.2SG	DEF:SG.F	bicicycle:SG.F			

'It is not like we live in the Netherlands, where you take the bicycle with four degrees below zero.'

b) Cases in which the relative clause was not fully realized, in that the speaker introduced the subordinator (i.e. the relativizing element) but then, the main verb is not produced and, thus, it was not possible to identify univocally the grammatical relation conveyed by the relativized element.
c) Occurrences realized by the interviewer (and not by interviewee).

This process has resulted in a datafile composed of 2.898 sentences that were manually annotated according to the following features.

First, the linguistic element employed was considered, in order to allow the discussion of their sociolinguistic characterization.

a) Relativizing element:
   i. ART. + *quale* ('which');
   ii. *cui* ('which');
   iii. *che* ('that');
   iv. *dove* ('where').

The sociolinguistic standardness of the occurrence was also annotated, using as a reference the Italian grammar authored by Serianni (2010 [1989]).

b) Sociolinguistic standardness:
   i. standard;
   ii. sub-standard.

Each occurrence was also tagged according to the strategy employed, adopting the taxonomy presented in § 1.

c) Strategy (strategy):
   i. relative pronoun;
   ii. gap;
   iii. pronoun retention;
   iv. double encoding.

Then, other linguistic features, semantic and syntactic in nature, that have traditionally been considered relevant in explaining variation and relativizing strategies were taken into account.

The grammatical relation that linked the antecedent with the relative clause was annotated, in order to verify with which strategies they were relativized. Nominative and accusative have been merged, given that they exhibit very little variability (see below) and can be relativized with the same strategies, since, as already mentioned, Italian does not have any dedicated preposition to mark nominative and accusative. Dative is expressed by the means of *a* ('to'), while genitive by *di* ('of') both followed by ART. + *quale* or *cui* ('which'). Locatives show a more heterogeneous behavior: several prepositions (followed by ART. + *quale* or *cui* 'which') can be employed, depending on the configuration of the described event, and *dove* 'where' alone can be selected. Furthermore, even though it is not traditionally considered a grammatical relation, we added the temporal value. This value is expressed by the means of a preposition (typically *in* 'in') followed by *quale* or *cui* ('which') or, differently from other non-nominative/accusative grammatical relations, by *che* ('that'). Given this latter structural possibility, we decided to control its behavior separately from other oblique relations.

For the sake of brevity, in the rest of the paper I will refer to the relations from ii. to vi. in d) as *obliques*; however, here *oblique* is to be understood as 'grammatical relations that *can* be relativized by the means of a preposition'. This label, basically, excludes only nominatives and accusatives, given that Italian does not have prepositions that express these grammatical relations.

d) Grammatical relation:
   i. nominative and accusative;
   ii. dative;
   iii. genitive;
   iv. locative;
   v. temporal;
   vi. other.

Then, all the occurrences were coded considering if the relativized element was an argument or an adjunct, in order to verify if the bond with the verb had a relevance in selecting the relativization strategy. As is well known, most of the arguments are nominative, accusative or dative but they can include also locative when a motion or a stative verb is involved.

e) Argument structure:
    i. argument;
    ii. adjunct.

Furthermore, all oblique relative clauses were annotated according to their semantics, i.e., it was tagged whether they were restrictive or non-restrictive, considering that in the former case they are considered to be more syntactically integrated within the sentence. As is well known, restrictive relative clauses allow for the identification of a referent among a set of possible referents, while non-restrictive relative clauses provide additional information about a referent.

f) Semantics:
    i. restrictive;
    ii. non-restrictive.

In order to distinguish the two categories, sentence negation was adopted as main criterion. As discussed by Cristofaro (2005: 195-196), negating a sentence containing a restrictive relative does not negate the content of the relative itself, as in (14a), while more interpretations are allowed when negating a sentence containing a non-restrictive relative clause, as in (14b).

(14) adapted from Cristofaro (2005: 195)
a. *The man [who is sitting in that office] is a psychologist.*
→ It is not true that he is a psychologist.

b. *They went to a number of Bach concerts, [for which they had booked tickets several months in advance].*
→ It is not true that they went to a number of Bach concerts; it is not true that they had booked tickets several months in advance; it is not true that they went to a number of Bach concerts, neither that they had booked tickets for them several months in advance.

Finally, other two linguistic parameters were annotated, in order to verify if they could play a role in the selection of the relativization strategy. First, I considered the target prepositions to verify if their diverse frequencies had consequences on the employed strategy. Then, I took into account the definiteness of the antecedent to verify whether a greater degree of accessibility favors the selection of more explicit syntactic strategies.

g) Preposition:
   i. *a*, ‘to’;
   ii. *con*, ‘with’;
   iii. *da*, ‘from’;
   iv. *di*, ‘of’;
   v. *fra/tra*, ‘between’ or ‘among’;
   vi. *in*, ‘in’;
   vii. *per*, ‘for’;
   viii. *su*, ‘on’;
   ix. *riguardo (a)*, ‘about’.

h) Definiteness of the antecedent:
   i. definite;
   ii. indefinite.

The main objective will be to discuss whether speakers with different social characterization use structurally different strategies for relativization. It will be considered whether and how different linguistic factors have relevance in the selection of different relativization strategies.

3. Discussion

After a brief overview over the frequencies of relative clauses in the two sub-corpora (H and L), the behavior of nominative/accusative and obliques will be discussed.

In table (3) are reported the absolute values of relative clauses in the two sub-corpora, taking into account their grammatical relation. Here and in the following tables, percentage values are displayed in brackets.

	Nom and Acc	**Obliques**	**Tot.**
H	1.445 (85,91%)	237 (14,09%)	1.682 (100%)
L	1.006 (82,73%)	210 (17,27%)	1.216 (100%)
			2.898

Table 3: Distribution: grammatical relations.

The first thing that can be noted is that relative clauses are more frequent in the productions of highly educated speakers. This is shown by the absolute values (1.682

vs. 1.216) and it is confirmed by the relative frequencies[5], which are 9,93 in H and 7,30 in L.

If we consider the distribution of the relative clauses in the two sub-corpora between the 2 types of grammatical relations, we note that the values are similar, even if some differences can be high-lightened. The vast majority of occurrences involve nominatives and accusatives, while all other cases are relativized more sporadically. However, speakers with higher educational achievements, proportionally, relativize nominatives and accusatives more frequently than the others (85,9% vs. 82,7%); and, specularly, speakers of the L corpus, proportionally, relativize obliques more often (17,3% vs. 14,1%). An analogous result has been observed comparing formal and informal spoken productions of Italian and in other languages (see Ballarè & Larrivée 2021); one could hypothesize that in lower productions speakers prefer to employ strategies different from relative clauses to modify a nominal head (such as the repetition of the nominal head itself) but further studies are needed.

Globally, the distribution is statistically significant at $p < 0,05$ (Fisher exact test statistic value is 0,0218).

3.1. Nominative and accusative

In this section the focus is on the relativization of nominative and accusative; in table (4) there are displayed the strategies selected in the two subcorpora. No cases of relative pronoun (i.e. ART. + *quale* 'which' and inflected variants) are attested and double encoding is not one of the options given that in Italian there is no case marking for nominative and accusative.

	Gap	**Resumptive pr.**	**Tot.**
H	1.438 (99,52%)	7 (0,48%)	1.445 (100%)
L	997 (99,11%)	9 (0,89%)	1.006 (100%)
			2.451

Table 4: Distribution: strategies (nominative/accusative).

[5] (number of occurrences / number of tokens of the sub-corpus)*1000.

The gap strategy is the one selected almost categorically. In the productions of highly educated speakers, it involves *che* ('that') in all the cases but 2, in which one speaker relativizes two nominatives selecting *dove* ('where'), as exemplified in (15). In L, an analogous situation is observed: *che* ('that') is selected in 994 cases over 997 and there are 3 occurrences of *dove* ('where') to relativize a nominative, as in (16). It is worth noting that in all the 5 cases in which *dove* ('where') is involved, the nominal antecedent is a location -as in (16)- or it is a derived form of a spatial noun, as in (15) where *meridionale* ('southerner') derives from *meridione* ('south').

(15) KIParla, PTB019

con	*il*	*meridionale*	*dove abitava*	*in via*	*Montenero*
with	DEF:SG.M	southerner:SG.M	REL	live:PST.3SG	in street	Montenero

'With the southerner who lived in Montenero street.'

(16) KIParla, PTA005

poi	*hai*	*il*	*bar*	*del*	*cinese*
then	have:PRS.2SG	DEF:SG.M	bar:SG.M	of. DEF:SG.M	chinese:SG.M
dove	*però*	*ha*	*una*	*sua*	*clientela*
REL	but	have:PRS.3SG	INDEF:SG.F	GEN.3.SG.F	clientele:SG.F

'Then you have the Chinese's bar, that has its clientele.'

If we consider the data, we can see that speakers with different educational achievements behave in a homogeneous way in informal productions and there are no significant differences[6]. This is true for the adopted strategies and selected linguistic items. That is to say that relativization strategies of nominative and accusative in informal spoken Italian are uniform regardless of the social characterization of the speakers. In fact, ART. + *quale* ('which') is completely absent and the employ of *che* ('that') is almost categorical. There are globally only 21 sub-standard occurrences out of 2.451, consisting of the employ of *dove* ('where') to relativize subjects (5 occurrences) and the co-occurrence of a pronoun with *che* ('that') (16 occurrences); these last occurrences always involved the

[6] The Fisher exact test statistic value is 0,3076 and the result is thus not significant at $p < 0,05$.

relativization of an accusative and the employ of a clitic pronoun, except in one case, reported in (17).

(17) KIParla, PTB002

Mi	*son*	*fermato*	*tante di quelle volte*			*da*
REFL.1SG	AUX.1SG	stop:PST.PTCP	many.times			to
questo	*mio*	*amico*	*che*	*lui*	*tante*	*volte*
DEM.SG.M	POSS.1SG	friend:SG.M	REL	SUBJ.3SG	many	times
usciva	*con*	*la*[7]				
go.out:PST.3SG	with	DEF:SG.F				

'I stopped many times at this friend of mine that used to go out with (her).'

3.2. Obliques

3.2.1 Distributions

As mentioned, the relativization of the obliques is where greater variability is expected. First, let us consider the distribution of non-standard realizations in the two sub-corpora presented in table (5).

	Standard	**Sub-standard**	**Tot.**
H	198 (83,54%)	39 (16,45%)	237 (100%)
L	115 (54,76%)	95 (45,24%)	210 (100%)
			447

Table 5: Distribution: standardness (obliques).

It is possible to observe how speakers in this case behave in diverse ways: while in H sub-standard realizations constitute only 16,46% of the occurrences, in L they are nearly half of the sample (45,24%). The distribution is statistically significant at $p < 0,01$ (Fisher exact test statistic value is $< 0,000001$).

[7] Unfortunately, the only example in which this strategy appears is a case of unconcluded utterance. Thus, it is not possible to complete the prepositional phrase. The presence of the definite feminine article (*la*) may lead us to think that the speaker wanted to mention a female person.

It is important to say that the non-standardness of the occurrences may be linked to the relativizing element or, more rarely, the selected preposition. In the rest of the section, the issue will be addressed more in depth.

Let us consider the structural strategies employed in the two sub-corpora in the relativization of the obliques reported in table (6).

	Rel. pron.	**Gap**	**Res. pron.**	**Double enc.**	**Tot.**
H	98 (41,35%)	125 (52,74%)	10 (4,22%)	4 (1,69%)	237 (100%)
L	19 (9,05%)	181 (86,19%)	10 (4,76%)	0 (0%)	210 (100%)
					447

Table 6: Distribution: strategies (obliques).

Overall, looking at the distribution of different relativization strategies in the two subcorpora, we can see macroscopic differences. In both cases, the gap strategy is the most frequently used: however, while in subcorpus H it is employed in just over half of the cases (52,74%), in subcorpus L it exceeds 86%. The second most frequently used strategy is the one involving a relative pronoun; again, however, the frequency values are very different: in H it exceeds 40% while in L it does not reach 10%. The remaining structures, i.e. resumptive pronoun and double encoding, are much rarer; interestingly, the double encoding (i.e. the double expression of the grammatical relation) is only attested in the productions of speakers with higher educational achievements (see Berretta 1993: 232). One example of resumptive pronoun strategy employed by a speaker with lower educational achievements is provided in (18).

(18) KIParla, PTB009

Tuo	*papà*	*e*	*l'*	*Elsa*	*che*	*la*	*nonna*
POSS.2SG	father.SG.M	and	DEF.SG.F	Elsa	REL	DEF.SG.F	grandmother.SG.F
Lidia	*gli*	*insegnava*	*la*	*matematica*			
Lidia	DAT.3PL	teach:PST.3SG	DEF.SG.F	mathemathics.SG.F			

'Elsa and your father, to whom grandmother Lidia taught mathematics.'

Speakers with low educational achievements prefer the only structure that does not involve case marking (i.e. gap); strategies involving a preposition or a clitic pronoun, overall, do not reach 14% of occurrences. Higher-educated speakers, on the other

hand, have more diverse behavior: although the gap strategy is the one employed most frequently, the others (i.e. relative pronoun, resumptive pronoun and double encoding) exceed 47%. A more detailed analysis of these differences will be addressed in the next section.

Before discussing the differences in terms of overt case-marking of the relativized item, it may be useful to look at the linguistic items selected by speakers with different educational achievements, that are reported in table 7. The topic, of course, ties in with the previous one since the different structures cannot be expressed by all the relativizing elements. In fact, one can have a case of relative pronoun or double encoding only with ART. + *quale* ('which') and *cui* ('which') -both preceded by a preposition-, while one can have gap and resumptive pronoun only with *che* ('that') and *dove* ('where')[8].

	ART. + *quale*	***Cui***	***Dove***	***Che***	**Tot.**
H	2 (0,84%)	100 (42,19%)	106 (44,73%)	29 (12,24%)	237 (100%)
L	1 (0,48%)	18 (8,57%)	84 (40,00%)	107 (50,95%)	210 (100%)
					447

Table 7: Distribution: relativizing element (obliques).

In line with what was observed for nominative and accusative, in this case, occurrences of the ART. + *quale* ('which') are rare (3 in total) in both H and L. However, we can note at least two major differences. The first is that *cui* ('which') is much more frequent in H (42,19% vs. 8,57%), and is the form selected for the expression of relative pronouns and double encodings. The second is that speakers in H prefer *dove* ('where') over *che* ('that') for the realization of gap strategy (44,73% vs. 12,24%); specularly, speakers in H use *che* ('that') more frequently than *dove* ('where'); *che* ('that') alone, in fact, is employed to relativize more than half of the obliques in the sub-corpus.

3.2.2 Explaining variation

Because of what was observed in the previous paragraph, it is of interest to discuss the differences in speakers' behavior by distinguishing between relativization

[8] Please note that no occurrences of PREP. + *dove* ('where') are attested in the corpus.

strategies that exhibit case marking (i.e. relative pronoun, pronoun retention and double encoding) and those that do not (gap).

In the literature, the topic has been approached in terms of *explicitness*, that is, how explicitly the strategy encodes the role of the antecedent (Comrie 1989: 163). Explicitness is described as gradual but, in our case, also because of the rather small dimension of the dataset, the parameter is treated as binary. As diverse as they are in terms of both structure and standardness, what is of interest here is the need to divide the strategies between those that involve an element, be it a clitic pronoun or be it a preposition, to make explicit the grammatical relation between the antecedent and subordinate clause and the others.

In this section, I discuss the results of a statistical analysis conducted by associating two values, i.e. case marked vs. non-case marked, to the scrutinized variable. The factors considered are (see § 2.2):

1) educational achievements of the speaker;
2) grammatical relation;
3) argument structure;
4) preposition;
5) definiteness of the antecedent.

The data will be analyzed adopting a conditional inference tree and a random forest (Tagliamonte & Baayen 2012; Levshina 2015), which are indicated when the dataset is unbalanced and rather small. A conditional inference tree is a decision tree used to model relationships between a target variable and more predictor variables. They use statistical tests to decide where to split the data in homogeneous sub-sets: the process involves selecting a predictor variable that has the strongest association with the target, then partitioning the data based on thresholds in that predictor. A random forest builds multiple conditional inference trees and combines their outputs to improve predictive accuracy. The result is a ranking of the selected parameters according to their importance.

The conditional inference tree is shown in figure (1). The C index is 0,83 and thus the model offers an excellent discrimination (Hosmer & Lemeshow 2000: 162).

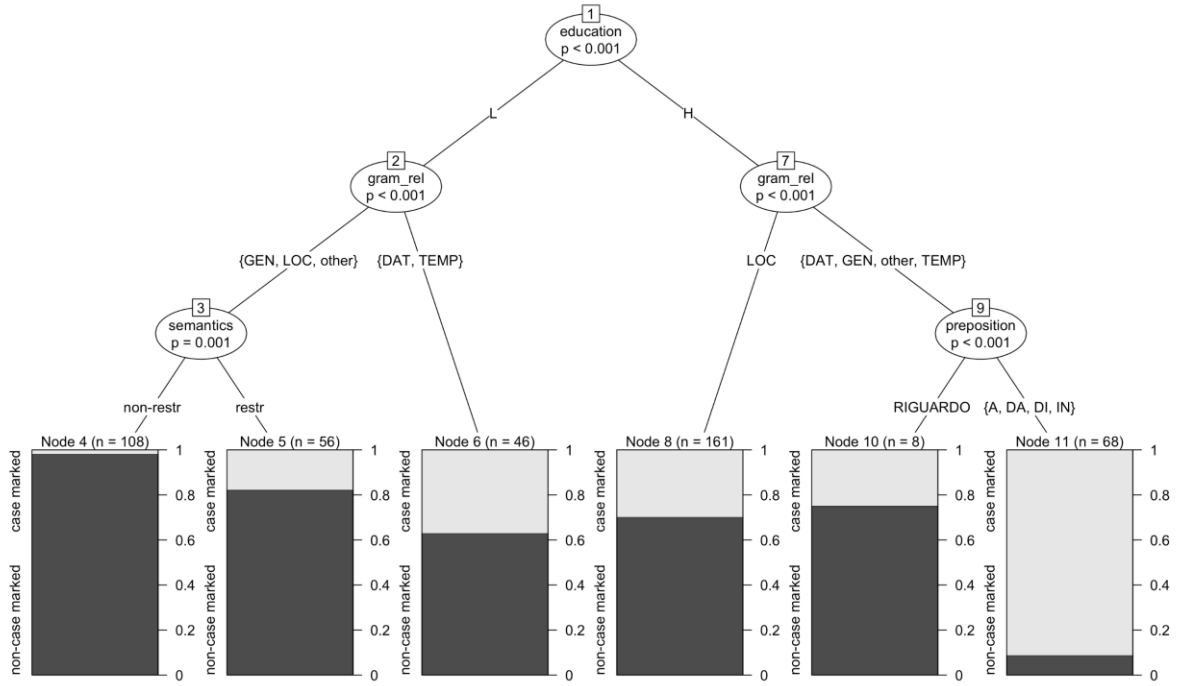


Figure 1: Conditional inference tree.

The first parameter that is of interest in creating homogeneous sub-part of the dataset is the social one, i.e. the educational achievement of the speakers (node 1). The behavior of speakers with lower educational achievements is represented on the left; the one of speakers with higher educational achievements, on the right. Overall, as already noted above, constructions with overt case marking are more frequent in the speech of subcorpus H, that is to say that speakers with higher educational achievements prefer using more complex structures in order to overtly express the grammatical relation relativized by the subordinate clause.

The second parameter relevant in both L and H is the grammatical relation of the relativized element. However, the partitioning of the data is done by grouping the values of the variable differently. In node 2, the division occurs between genitive, locative and *other* vs. dative and temporal. The first group is most often realized by the gap strategy. The locative in Italian has the dedicated form *dove* ('where') and this may be why the gap strategy is preferred. In *other* are placed relations that are rarer and sometimes realized through prepositions less frequent than the others and/or improper (see below). Surprising, however, is the placement of the genitive since it has no dedicated relativizing element, but it is still more frequently realized without the overt case marking. In the second group, i.e. dative and temporal, the gap strategy is still the most frequent but, proportionally, the percentage of case-marked structures

is higher. That is to say that these two relations, compared with the others, are more frequently expressed by non-gap strategy. This is explainable for dative, while it is less explicable for temporals, which, at least according to some grammars, in Italian can be relativized using *che* ('that').

At this point, only in the right portion of the tree and only for the grammatical relations of genitive, locative and *other*, the semantics of the relative (node 3) gain relevance. Not surprisingly, in non-restrictive relatives, thus less syntactically integrated in the sentence, non-case marked strategies (node 4) are more frequent than in restrictive ones (node 5).

Let us now consider the behavior of speakers in subcorpus H, where two linguistic factors come into play. The first, as already mentioned, is the grammatical relation (node 7). If the clause relativizes a locative relation, then it is expressed by a gap strategy in most cases. The difference in behavior between this function and the others is easily explained by the aforementioned presence of the dedicated form (node 8). All other relations are more frequently realized with a case marked strategy even if another factor acquires relevance: the target preposition of the subordinate clause (node 9). In fact, speakers do not use a case marking more frequently when an improper preposition, i.e. *riguardo* ('about'), is employed (node 10). Even though the number of occurrences is rather low (8), two things are worth saying. The first is that *riguardo* ('about') is not polyfunctional and much less frequent in the corpus than the other prepositions and therefore probably less easily retrievable by the speaker. In table (8) I show the normalized frequencies[9] of the different prepositions within the ParlaTO corpus.

Preposition	**Normalized frequency**
A 'to'	1,4%
Da 'from'	0,5%
Di 'of'	1,6%
Con 'with'	0,4%
In 'in'	1,1%
Per 'for'	0,7%
Su 'on'	0,1%
Tra/fra 'between' or 'among'	0,1%
Riguardo (a) 'about'	0,0027%

Table 8: Prepositions' frequencies.

[9]Percentage values are reported to the first decimal place, except for *riguardo* ('about') since its value is very low.

The second thing that can be worth mentioning is that the aboutness value is close to the one of the subject from an informative point of view and, thus, this could be one of the reason why it triggers the selection of a gap strategy, which is typically used to relativize subjects; furthermore, it has been observed that in Italian this values is often realized by the means of *dove* ('where') - see Ballarè & Inglese (2022).

If another preposition is involved, speakers in H select almost categorically a case marked strategy (node 11).

We can now consider the importance of factors in explaining the selection of the relativization strategy in the whole dataset. Figure 2 shows the random forest ranking; its C index is 8.6 (excellent discrimination, Hosmer & Lemeshow 2000: 162) and below are the numerical values obtained from the analysis:

- gram_rel: 0,058;
- education: 0,046;
- semantics: 0,004;
- definiteness: 0,003;
- preposition: 0,001;
- argument structure: 0,001.

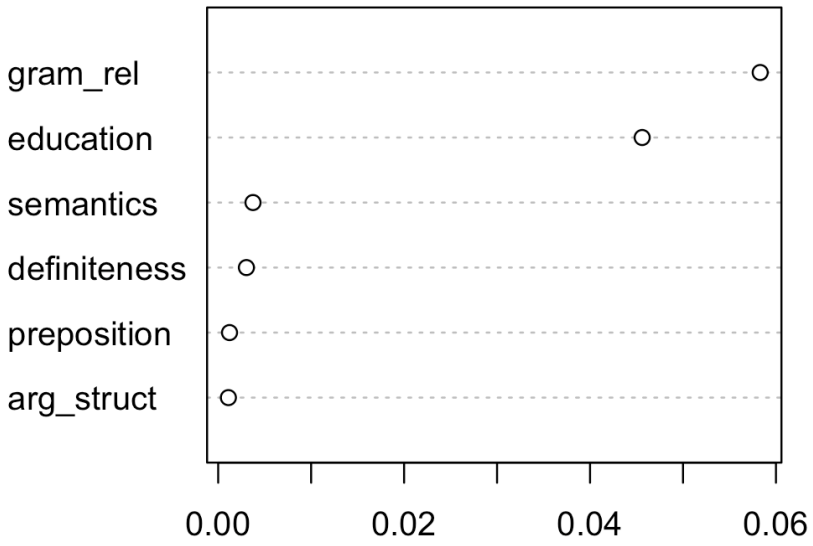


Figure 2: Random forest.

The first two parameters (i.e., grammatical relation and educational achievements) have importance in selecting the relativization strategy; the other four (i.e., semantics, antecedent definiteness, target preposition and argument structure) do not. As can be seen, the random forest indicates the importance of the parameter but not the

direction in which it acts; therefore, below I discuss the distributions associated with the prominent values.

In the first instance, I take into account the values associated with the grammatical relation, shown in table 9. Percentages are calculated based on total column value. The chi-square statistic is 62.087; the *p*-value is $< 0{,}00001$, and thus the result is significant at $p < 0{,}01$.

	dat	**gen**	**temp**	***other***	**loc**
Case marked	13 (72,22%)	13 (61,90%)	35 (59,32%)	25 (35,21%)	55 (19,78%)
Non-case marked	5 (27,78%)	8 (38,10%)	24 (40,68%)	46 (64,79%)	223 (80,22%)
	18 (100%)	21 (100%)	59 (100%)	71 (100%)	278(100%)

Table 9: Strategy and grammatical relation.

Grammatical relations are entered from left to right by those most frequently having overt case marking and decreasing.

Before moving to the analysis, it is important to note that the absolute values of the first two columns are rather low and thus any generalization requires due caution.

Dative and genitive are expressed in the vast majority of cases through the use of a case mark. The temporal and locative values, i.e. the only ones that, according to grammars, admit both strategies, show different behaviors. Temporal values are expressed by employing both available strategies (with a preference for overt case marking), while locative ones are realized more frequently (80,22% of cases) through a non-case marked strategy.

Even if grammatical relations and not semantic roles were tagged, the results can be discussed in relation to a well-known typological generalization. Comrie, in fact, states that “the more difficult a position is to relativize, the more explicit indication is given of what position is being relativized, to facilitate recovery of this information” (Comrie 1989: 163). With “difficulty of relativization”, Comrie is referring to the well-known accessibility hierarchy (Keenan & Comrie 1977: 66), which “is basically the degree of salience of the participant in the relative clause event” (Croft 2022: 604); the further to the left the position, the easier the relativization.

SUBJECT > DIRECT OBJ. > INDIRECT OBJ. > OBLIQUE > GENITIVE > OBJ. OF COMPARISON

Following Comrie (1989: 163), this means that "for instance, where the choice is between a pronoun-retention and a gap relative clause, it is nearly always the case that the pronoun-retention type is used lower down the accessibility hierarchy […], while the gap strategy is used higher up".

Considering the Italian data, the generalization remains valid for subject and direct object (approximated to nominative and accusative, with some exceptions) since, as discussed, they are relativized almost categorically with the least explicit structure, namely *che* ('that'). However, the position of the genitive is problematic, since it precedes all the obliques (locative, temporals and *other*) and the indirect object (approximable, again, not without exceptions, to the dative).

The order found, however, albeit with some differences, is reminiscent of another typological hierarchy, namely the inflectional case hierarchy proposed by Blake (2001), reported below.

NOM ACC/ERG GEN DAT LOC ABL/INST *others*

At is known, the hierarchy is to be interpreted as follows: "if a language has a case listed on the hierarchy, it will usually have at least one case from each position to the left" (Blake 2001: 156). Between the scrutinized relativization strategies and the case hierarchy, there are some substantial differences and limits. First, Blake considers only morphological case systems while, in our case, I am referring to a language that, in most cases, expresses grammatical relations by prepositions. Moreover, inflectional case, typically, act on a word level while I am considering strategies that are at work on a syntactic one. The hierarchy also has some limits, since there are several exceptions to it, also given by cases of syncretism (see Baerman et al. 2005 inter al.).

However, when comparing the inflectional case hierarchy and the result of our analysis (disregarding nominative and accusative), it is possible to note the presence first of genitive and dative in a rather high position on the scale, and then the other values. As already mentioned, locatives and temporals in Italian exhibit a different structural behavior compared to other obliques. Although with some differences, what is interesting to note here is that, when possible, speakers proportionally more frequently employ an overt case marking relativization strategy in an order that reminds the one according to which the languages of the world have a dedicated case mark.

The explanation, thus, is not to be found in the accessibility of the participant but more on a syntactic level. The first two grammatical relations, in fact, have a deeper

bond to the main clause with respect to the others: typically, in fact genitive modifies a nominal element, while the dative is a verb argument. The other relations, on the other hand, are, at least prototypically, less bonded and behave more often as adjuncts of the main clause. Syntax, however, also given the non-importance of the argument/adjunct parameter does not tell the whole story.

In fact, it is worth noting that, on a semantic level, dative and genitive, differently from the other grammatical relations, are often linked with animacy[10]. One could argue that, given the higher saliency of animate referent and in order to avoid ambiguity, speakers make explicit the grammatical relation between the antecedent and the relative clause, by selecting a more structurally complex strategy.

The second parameter that shows importance is educational achievements. The distribution is given in table (10); the Fisher exact test statistic value is $< 0{,}00001$ and thus the result is significant at $p < 0{,}01$.

	H	**L**
Case marked	112 (47,26%)	29 (13,81%)
Non-case marked	125 (52,74%)	181 (86,19%)
	237 (100%)	210 (100%)

Table 10: Strategy and educational achievement.

As discussed in the previous section, speakers with lower educational achievements prefer to adopt the less complex strategy (86,19%), while those with higher educational achievements show values that hover around 50% for both types of strategies at their disposal; thus, proportionally, they use the more complex strategies more often and explicate the syntactic relation between the antecedent and subordinate clause.

Even though the statistical model considers all other parameters to be not important in explaining how speakers select different relativization strategies, it may be useful to note that the distribution of two of these, i.e., semantics and argument structure (cfr. Fiorentino 1999: 167), turns out to be statistically significant (with $p < 0{,}01$ and Fisher exact test value of 0,0007 and 0,0036, respectively). The apparent contradiction is actually easily resolved by clarifying the differences statistical

[10] The relevance of animacy in explaining cases of *pronoun retention* has been noted also by Berretta (1993: 232-233) and Fiorentino (1999: 104).

significance and model importance. In fact, statistical significance measures whether a variable's distribution differs from random chance, often in isolation, while random forest importance assesses a variable's contribution to prediction accuracy within the context of all other variables.

In order to understand more clearly the behavior of the relativizing structures, the distributions of these last two parameters are shown in Tables 11 and 12.

	restr	**non-restr**
Case marked	75 (40,54%)	66 (25,19%)
Non-case marked	110 (59,46%)	196 (74,81%)
	185 (100%)	262 (100%)

Table 11: Strategy and semantics.

	arg	**adj**
Case marked	49 (42,61%)	92 (27,71%)
Non-case marked	66 (57,39%)	240 (72,29%)
	115 (100%)	332 (100%)

Table 12: Strategy and argument structure.

The distributions show quite clearly that when the semantic (i.e. "restrictive") or syntactic (i.e. "argument") connection between the relative subordinate and the main sentence is stronger, speakers more frequently select case-marked strategies (40,54% vs. 25,19% and 42,61% vs. 27,71%, respectively).

4. Conclusive remarks

The results of the conducted analysis shed light on relativization strategies in the sub-standard area of spoken Italian.

In general, relative clauses are more frequently realized by speakers with higher educational achievements.

For the relativization of nominative and accusative, the behavior in the sub-standard area is uniform: all speakers, regardless of their social characterization, employ *che* ('that'). It is important to emphasize that, from a functional point of view, selecting *che* ('that') or ART. + *quale* ('which') has no substantial consequences given

that in Italian, in this domain, there are no case markers for nominative and accusative. The difference between these two linguistic elements is sociolinguistic in nature: in more controlled contexts, the use of ART. + *quale* ('which') remains frequent, perhaps precisely to mark the formality of the production. There are few occurrences in which a clitic pronoun is also involved, and it is interesting to note that, in terms of frequency, speakers in H and L behave in an analogous way by producing a similar number of sub-standard occurrences, regardless of their educational achievements.

Profound differences have been observed in the relativization of obliques, depending on the social characteristics of the speakers.

Not surprisingly, speakers with lower educational achievements produce more sub-standards relatives; what is of interest, however, is that while they produce them by simplifying the structure and over-extending the gap strategy, speakers with higher educational achievements realize sub-standard occurrences complexifying the structure and employing the double-encoding strategy.

We also notice differences in the selection of relativizing elements: speakers in H frequently use *cui* ('which') and prefer *dove* ('where') over *che* ('that'); speakers in L, on the other hand, use *che* ('that') significantly more frequently. This, of course, ties in with the relativized relations. Statistical analysis showed that there is a significant relation between educational achievements and the adoption of case marked or non-case marked relativization strategies, since highly educated speakers prefer to overtly mark the grammatical relation. Furthermore, the most important factor in selecting a strategy type is the grammatical relation. Genitive and dative, that are syntactically bonded to the main clause and that are the only oblique relations that can involve an animate referent, are the ones more frequently expresses by a case marked strategy.

From a sociolinguistic perspective, we can say that the homogeneity detected for nominative and accusative is *not* found in the obliques because speakers behave significantly differently. That is to say that, at least in our data and at least for the relativization of the obliques, speakers with lower educational achievement select different strategies compared to others. Speakers in H, on the other hand, show greater variability and have more relativization strategies at their disposal.

Studying thoroughly data of a single language from a sociolinguistic perspective allows for an accurate analysis, that also considers specific features of the scrutinized language itself; however, the study shows also that the adoption of typological categories allow us to go beyond them and tie the results to the bigger picture.

Abbreviations

1 = 1st person
2 = 2nd person
3 = 3rd person
ABL = ablative
ACC = accusative
ART = article
AUX = auxiliary
COMP = complementizer
DAT = dative
DEF = definite
DEM = demonstrative
ERG = ergative
F = feminine
GEN = genitive
INDEF = indefinite
INF = infinitive
INS = instrumental
LOC = locative
M = masculine
NEG = negation
NOM = nominative
OBJ = object
PL = plural
POSS = possessive
PRS = present
PST = past
PTCP = participle
REFL = reflexive
REL = relative
SG = singular
SUBJ = subject

References

Ackerman, Farrell & Irina Nikolaeva. 2013. *Descriptive typology and linguistic theory: a study in the morphosyntax of relative clauses*. Stanford: CSLI.

Alexiadou, Artemis, Paul Law, André Meinunger & Chris Wilder (eds.). 2000. *The syntax of relative clauses*. Philadelphia: John Benjamins.

Alfonzetti, Giovanna. 2022. *La relativa non standard. Italiano popolare o italiano parlato?*. Palermo: Centro di studi filologici e linguistici siciliani.

Alisova, Tatiana. 1965. Relative limitative e relative esplicative nell'italiano popolare. *Studi di filologia italiana XXIII*. 299-333.

Baerman, Matthew, Dunstan Brown & Greville G. Corbett. 2005. *The Syntax-Morphology Interface. A Study of Syncretism*. Cambridge: Cambridge University Press.

Ballarè, Silvia. 2024. L'italiano colloquiale. In Silvia Ballarè, Ilaria Fiorentini & Emanuele Miola (eds.), *Le varietà dell'italiano contemporaneo*, 81-98. Roma: Carocci.

Ballarè, Silvia & Pierre Larrivée. 2021. Register impacts syntax: scaling the accessibility hierarchy of relatives. *Italian journal of linguistics 33*. 3-22.

Ballarè, Silvia & Guglielmo Inglese. 2022. The development of locative relative markers from typology to sociolinguistics (and back). *Studies in language 46*. 220-257.

Benincà, Paola, Mair Parry & Diego Pescarini. 2016. The dialects of northern Italy. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford Guide to the Romance languages,* 185-205. Oxford: Oxford University Press.

Bernini, Giuliano. 1989. Tipologia delle frasi relative italiane e romanze. In Fabio Foresti, Elena Rizzi & Paola Benedini (eds.), *L'italiano tra le lingue romanze. Atti del XX Congresso*

Internazionale della Società di Linguistica Italiana (Bologna, 25-27 settembre 1986), 85-98. Roma: Bulzoni.

Berretta, Monica. 1988. Varietätenlinguistik des Italienischen / Linguistica delle varietà. In Günter Holtus, Michael Metzeltin & Christian Schmitt (eds.), *Lexicon der Romanistischen Linguistik. Italienisch, Korsisch, Sardisch*, 762-774. Berlin/New York: Mouton de Gruyter.

Berretta, Monica. 1993. Morfologia. In Alberto A. Sobrero (ed.), *Introduzione all'italiano contemporaneo* I, 193–245. Roma/Bari: Laterza.

Berruto, Gaetano. 1983. Italiano popolare e semplificazione linguistica. *Vox Romanica* 42. 38-79.

Berruto, Gaetano. 2012. *Sociolinguistica dell'italiano contemporaneo (seconda edizione)*. Roma: Carocci.

Berruto, Gaetano. 2014. Esiste ancora l'italiano popolare? Una rivisitazione. In Paul Danler & Christine Konecny (eds.), *Dall'architettura della lingua italiana all'architettura linguistica dell'Italia. Saggi in omaggio a Heidi Siller-Runggaldier*, 277-190. Frankfurt am Main: Lang.

Blake, Barry J. 2001. *Case. Second edition*. Cambridge: Cambridge University Press.

Cerruti, Massimo. 2016. Costruzioni relative in italiano popolare. In Federica Guerini (ed.), *Italiano e dialetto bresciano in racconti di partigiani*, 77–116. Roma: Aracne.

Cerruti, Massimo. 2017. Changes from below, changes from above. Relative constructions in contemporary Italian. In Massimo Cerruti, Claudia Crocco & Stefania Marzo (eds.), *Towards a new standard. Theoretical and empirical studies on the restandardization of Italian*, 32-61. Berlin/New York: Mouton de Gruyter.

Cerruti, Massimo & Silvia Ballarè. 2021. ParlaTO: corpus del parlato di Torino. *Bollettino dell'Atlante linguistico Italiano 44*. 13-38.

Cinque, Guglielmo. 2020. *The syntax of relative clauses: a unified analysis*. Cambridge: Cambridge University Press.

Comrie, Bernard. 1989. *Language universals and linguistic typology. Second edition*. Chicago: University of Chicago Press.

Comrie, Bernard & Tania Kuteva. 2013a. Relativization on Subjects. In Matthew Dryer & Martin Haspelmath (eds.), *World atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/122. (Accessed 2024-10-25).

Comrie, Bernard & Tania Kuteva. 2013b. Relativization on Obliques. In Matthew Dryer & Martin Haspelmath (eds.), *World atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/123. (Accessed 2024-10-25).

Cortelazzo, Manlio. 1972. *Avviamento critica allo studio della dialettologia italiana, III, Lineamenti di italiano popolare.* Pisa: Pacini.

Cristofaro, Sonia. 2005. *Subordination*. Oxford: Oxford academic.

Cristofaro, Sonia & Anna Giacalone Ramat. 2007. Relativization strategies in the languages of Europe. In Paolo Ramat & Elisa Roma (eds), *Europe and the Mediterranean as linguistic areas: Convergences from a historical and typological perspective*, 63–93. Amsterdam/Philadelphia: John Benjamins.

Croft, William. 2022. *Morphosyntax. Constructions of the World's Languages.* Cambridge: Cambridge University Press.

De Mauro, Tullio. 1970. Per lo studio dell'italiano popolare unitario. In Annabella Rossi (ed.), *Lettere da una tarantata*, 43-57. Bari: De Donato.

Fiorentino, Giuliana. 1999. *Relativa debole. Sintassi, uso, storia in italiano*. Milano: Franco Angeli.

Henderey, Rachel. 2012. *Relative clauses in time and space: a case study in the methods of diachronic typology*. Amsterdam/Philadelphia: John Benjamins.

Hosmer, David W. & Stanley Lemeshow. 2000. *Applied logistic regression*. New York: Wiley.

Inglese, Guglielmo & Silvia Ballarè. 2023. Analyzing language variation: where sociolinguistics and linguistic typology meet. In Silvia Ballarè & Guglielmo Inglese (eds.), *Sociolinguistic and typological perspectives on language variation*, 1-28. Berlin: Mouton De Gruyter.

itTenTen20 corpus, available on https://www.sketchengine.eu/ittenten-italian-corpus/

Keenan, Edward L. & Bernard Comrie. 1977. Non phrase accessibility and universal grammar. *Linguistic inquiry 8*. 63-99.

Kidd, Evan (ed.). 2011. *The acquisition of relative clauses: processing, typology and function.* Amsterdam/Philadelphia: John Benjamins.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh & Joeal Scherzer (eds.), *Language in use: readings in sociolinguistics*, 28-54. Englewood Cliffs (NJ): Prentice Hall.

Ledgeway, Adam. 2016. The dialects of southern Italy. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford Guide to the Romance languages,* 246-269. Oxford: Oxford University Press.

Ledgeway, Adam & Martin Maiden (eds.). 2016. *The Oxford Guide to the Romance languages.* Oxford: Oxford University Press.

Lepschy, Giulio. 2002. Popular Italian: Fact or Fiction?. In Giulio C. Lepschy (ed.), *Mother Tongues and Other Reflections on the Italian Language*, 49-69. Toronto: University of Toronto Press.

Levshina, Natalia. 2015. *How to do Linguistics with R*. Amsterdam: John Benjamins.

Mauri, Caterina, Silvia Ballarè, Eugenio Goria, Massimo Cerruti & Francesco Suriano. 2019. KIParla corpus: a new resource for spoken Italian. In Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds.). *Proceedings of the 6th Italian conference on Computational Linguistics CLiC-it*, Torino, Accademia University Press.

Murelli, Adriano. 2011. *Relative constructions in European non-standard varieties*. Berlin/New York: Mouton de Gruyter.

Renzi, Lorenzo. 2000. Le tendenze dell'italiano contemporaneo. Note sul cambiamento linguistico nel breve periodo. *Studi di lessicografia italiana XVII*. 279-319.

Renzi, Lorenzo. 2012. *Come cambia la lingua. L'italiano in movimento*. Bologna: Il Mulino.

Serianni, Luca. 2010 [1989]. *Grammatica italiana: italiano comune e lingua letteraria. Suoni, forme, costrutti*. Torino: UTET.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as case study for statistical practice. *Language variation and change 24(2)*. 135-178.

CONTACT

silvia.ballare@unibo.it

Linguistic Typology at the Crossroads 4-2 (2024): 158-219 ISSN 2785-0943

Are asymmetries in imperative negation based in usage?

DANIEL VAN OLMEN
LANCASTER UNIVERSITY

Submitted: 29/04/2024 Revised version: 1/08/2024
Accepted: 7/08/2024 Published: 23/01/2025

Abstract

This article extends the study of (a)symmetries in negation to the domain of (negative) imperatives. It examines a balanced sample of the world's languages for distinctions in tense, direction/location and intersubjectivity and observes that, like with asymmetry in standard negation, they are often neutralized from positive to negative but not vice versa. Intersubjective marking is found to be somewhat exceptional in that the opposite situation does occasionally occur. The article also tests whether and confirms that these asymmetries are grounded in usage patterns, with a corpus investigation of English and Dutch (negative) imperatives. It proposes negation's discourse presuppositionality, which has been argued to account for neutralization in standard negation, as an explanation for most but not all of these typological and usage-based results in imperative negation too. It nevertheless makes a case for other, more imperative-specific motivations as well.

Keywords: asymmetry; Dutch; English; (negative) imperative; usage.

1. Introduction

The notion of asymmetry at the heart of this article comes from Miestamo's (2005) typological study of standard negation. He characterizes (a)symmetry in this way: domain *f(*x*)* is symmetric if its grammatical structures differ from those of x only in the presence of the *f()* marking; if there are more differences, *f(*x*)* is asymmetric (Miestamo 2005: 51–52). For imperative negation, symmetry is thus the situation

where the negation – i.e. *f()* – of the imperative – i.e. x – simply involves extra negative marking, as in Dutch in (1).

(1) Dutch (NLD; Germanic, Indo-European; personal knowledge)
ga (*niet*) *weg*
go.IMP NEG away
'(Don't) go away!'

Asymmetry can be constructional and/or paradigmatic (Miestamo 2005: 51–56). In Pite Saami, for example, all paradigmatic distinctions made in the imperative are also available in the negative imperative but the latter construction consists of a prohibitive auxiliary bearing the imperative marking and a non-finite "connegative" form of the lexical verb, as (2) shows.[1] We can say that the negative imperative in (2b) is constructionally asymmetric vis-à-vis its positive equivalent in (2a).

(2) Pite Saami (SJE; Saami, Uralic; Wilbur 2014: 152, 158, 180)
a. *dáhke-n dal d-a-v*
do-DU.IMP now DEM-DIST-ACC.SG
'You two do that now!'
b. *elle-n tsábme*
PROH-DU.IMP eat.CONNEG
'Don't you two eat!'

In Matsés, imperative negation does exhibit constructional symmetry: the imperative in (3a), marked by the absence of any inflection, and its negative counterpart in (3b) differ only in the prohibitive suffix *-enda*.

(3) Matsés (MCF; Panoan; Fleck 2003: 993)
a. *cun shubu-no nid*
1.GEN house-LOC go
'Go to my house!' (speaker *might* accompany addressee)

[1] The term "prohibitive" is sometimes used in place of or preferred to "negative imperative". We will stick to this second label and reserve the first one for markers dedicated to expressing 'don't!', like *elle* in (2b).

b. *cun* *shubu-no* *nid-enda*
1.GEN house-LOC go-PROH
'Don't go to my house!'

c. *cun* *shubu-no* *nid-ta*
1.GEN house-LOC go-N1.IMP
'Go to my house!' (speaker won't accompany addressee)

There is paradigmatic asymmetry, though. In the positive, the language has the option of adding *-ta*, as in (3c), to signal that the speaker will not join the addressee in the action. The imperative in (3a) leaves the speaker's involvement unspecified. Crucially, in the negative, this distinction is not available.

Applying this concept of (a)symmetry to standard negation has enabled Miestamo (2005) to uncover a range of recurring phenomena in the world's languages. He observes, for instance, that languages distinguishing realis and irrealis grammatically often feature additional irrealis marking, either compulsorily or optionally, in negative declarative verbal main clauses. Moreover, such sentences are never found to be realis-marked whilst their positive equivalents are irrealis-marked. This asymmetry is, in his view, motivated by the fact that, "semantically, negation belongs to the realm of the non-realized" (Miestamo 2005: 196; see Cristofaro 2012: 140–142, however, for examples of how irrealis arises in standard negation through diachronic processes unrelated to the domain of the non-realized). Another asymmetric phenomenon identified by Miestamo (2005) for standard negation is the frequent neutralization of positive tense-aspect-mood and person-number-gender distinctions in the negative. An example of his is Bagirmi. The positive construction in (4a) has a symmetric negative counterpart in (4b) but negation is incompatible with completive *ga* in (4c) and the aspectual distinction between (4a) and (4c) is therefore lost in (4b).

(4) Bagirmi (BMI; Bongo-Bagirmi, Central Sudanic; Stevenson 1969: 83, 91, 130)

a. *ma* *m-'de*
1SG 1SG-come
'I came.'

b. *ma* *m-'de* *li*
1SG 1SG-come NEG
'I didn't come.'

c. *ma* *m-‘de* *ga*
1SG 1SG-come COMPL
‘I have come.’

To account for this neutralization tendency, Miestamo (2005: 211) appeals to the idea of discourse presuppositionality: “Since negatives [e.g. *he didn’t break the rules*] typically occur in contexts where the corresponding affirmative [i.e. *he broke the rules*] is supposed or somehow present, many aspects of the negated content are known to the speakers, and there is less need to explicitly specify its different properties such as its temporal aspects or its participants.” This explanation centers around what it is basically an assumed discourse preference. As Miestamo et al. (2022: 135–136) argue, it would then have “conventionalized as grammatical constraints” in languages like Bagirmi but, importantly, one should/would expect its effects to “be present in all languages” – in patterns of usage to be precise, of grammatical as well as lexical expressions. In this regard, it is interesting to note Miestamo et al.’s (2024: 22–26) findings for declarative verbal main clauses in Korean, English and Finnish conversations: temporal adjuncts indeed occur less often in negative than in positive sentences, though only in the former two languages significantly so, and the same holds for the adjuncts of temporal position in particular (which locate a state of affairs in time, compared to those of duration, frequency or temporal relationship; see Hasselgård 2010: 204–206), though only in the latter two languages significantly so. Miestamo et al. (2024: 26–27) see these results as partial confirmation for the claim that tense neutralization is “motivated by the lower need for temporal specification in negatives”. Such bases in usage are what this article aims to examine for asymmetries not in standard negation but in imperative negation.

(Negative) imperatives have been studied and compared from a cross-linguistic perspective before (e.g. Xrakovskij 2001; Mauri & Sansò 2011; van der Auwera & Lejeune 2013; Van Olmen 2021). However, remarkably little attention has been paid to recurrent patterns of constructional or paradigmatic variation between the imperative and its negative counterpart. Even less research has looked at such patterns using the notion of (a)symmetry, despite Miestamo’s (2005: 238) call “to broaden the scope of the study [of (a)symmetries] into other areas of clausal negation, especially into non-declarative negation”. One of the exceptions is Miestamo & van der Auwera (2007). They consider just 30 languages, though, and they primarily seek to answer the question to what extent imperative negation exhibits the asymmetries known from

standard negation. For a more singular focus on imperatives versus negative imperatives, we can turn to Aikhenvald (2010: 165–197). She may not describe their (dis)similarities as (a)symmetries but she makes numerous observations that can quite easily be recast in such terms. She notes, for instance, that, "in many languages, … categories [relating to verbal action] are found in positive, but not in negative, imperatives" (Aikhenvald 2010: 181). Another way to formulate this observation would be to say that imperative negation frequently displays asymmetry of the paradigmatic neutralization type. Tucano in (5) is one of her examples and involves a tense distinction particular to (negative) imperatives across languages (Aikhenvald 2010: 129–131), i.e. between immediate compliance in (5a) and delayed compliance in (5b). The negative imperative in (5c) is said to correspond to both (5a) and (5b).[2]

(5) Tucano (TUO; Tucanoan; West 1980: 51; Aikhenvald 2010: 183)
   a. *ba'á-ya*
      eat-IMM.IMP
      'Eat now!'
   b. *ba'a-apa*
      eat-DEL.IMP
      'Eat later!'
   c. *ba'a-tikaya*
      eat-PROH
      'Don't eat!'

Aikhenvald (2010: 183) adds that languages may also retain the distinction between immediate and delayed compliance in negative imperatives. No indications of the frequency of (non-)neutralization are provided, though, and the (im)possibility of tense being marked only in the negative imperative is simply not discussed. It therefore seems warranted to have another, closer look at this asymmetry, which is precisely what the present study seeks to do. Immediate versus delayed compliance is, however, not the sole feature that merits revisiting. We will investigate two further types of distinctions: directional-locational ones (e.g. 'go and …!') and intersubjective ones (e.g. illocutionary force). There are, of course, many others that might be of

[2] We are following Aikhenvald's (2010: 183) analysis here. West (1980: 51) does mention the delayed negative imperative *ba'a-tí-cã'-apa* (eat-NEG-EMP-DEL.IMP) 'at a later point, don't eat!', alongside *ba'a-tí-cã'-ña* (eat-NEG-EMP-IMM.IMP) 'don't eat, now!'.

interest (e.g. number marking or (im)perfective aspect in (negative) imperatives). Our focus on the three types of distinctions just mentioned is motivated in part by space limitations. A more significant reason is that, in many languages, these distinctions are specific to or, put differently, made solely in (negative) imperatives but their (dis)similarities in the positive and the negative have only been explored cursorily (see Aikhenvald 2010: 133–138, 183–184, 189–190, 203–223).

In short, we want to examine in this article (i) whether imperative negation exhibits any systematic asymmetries in tense, directional-locational and intersubjective distinctions and, if so, how (often) they manifest themselves cross-linguistically and (ii) whether and in what way any such asymmetries can be accounted for by considering usage data. To answer (i), we will take a typological perspective in Section 2 and look at a balanced sample of 160 of the world's languages. To answer (ii), we will adopt a usage-based perspective in Section 3 and investigate corpus data of both English and Dutch. Section 4, finally, will contain our conclusions.

2. Typological perspective

This section will first discuss our sample (Section 2.1). Then, we will focus on the marking in imperative negation of tense (Section 2.2), direction and/or location (Section 2.3) and intersubjectivity (Section 2.4). An interim summary will be given at the end (Section 2.5).

2.1. Sample

For our typological study, we rely on a 160-language sample that follows Miestamo et al.'s (2016: 256–259) genus-macroarea sampling method with a predetermined sample size. This method produces a variety sample, which primarily serves to reveal as much diversity as possible in how the languages of the world convey some functional domain, like imperative negation. To be reliable, it should "represent all the world's linguistic groupings – areal, genealogical and other – as well as possible", since "connections between languages increase the possibility that they are similar to each other" (Miestamo et al. 2016: 235). If such representation is attained by eliminating potential biases in a consistent way, the variety sample may even be used to make claims about, for example, cross-linguistic frequency (Miestamo et al. 2016: 251–252).

To limit genealogical bias, the present method takes Dryer's (1989) concept of genus as its point of departure. Genera are linguistic groupings for which one can reconstruct a common ancestor that is normally between 3,500 and 4,000 years old (Dryer 1989: 267). A genus may belong to a bigger language family (e.g. Sinitic), make up an entire language family itself (e.g. Mayan) or be an isolate (e.g. Warao[3]). Starting from genera for a sample's genealogical classification has the benefit that, unlike many language families, they constitute groupings of languages that are quite generally accepted as related (Miestamo et al. 2016: 238–239). Dryer (2013) lists 521 such groupings for the world's languages and our sampling method stipulates that none of these genera can be represented by more than one language. In theory, the choice of language could/should be arbitrary but, in practice, it is obviously affected by the (un)availability of sufficient information. Lack of data has an impact on the selection of genera too. There is many a genus of which no language has been adequately documented (yet) and that cannot but be excluded from the sample. Moreover, such genera tend to be more common in some areas (e.g. Australia) than in other ones (e.g. Europe) (Miestamo et al. 2016: 250). The former would be underrepresented and the latter overrepresented in a sample that simply included a language from any genus with enough information. Some geographical stratification is therefore needed.

To mitigate areal and bibliographical bias, Miestamo et al. (2016: 256) draw on Dryer's (1992) six so-called macroareas, the more or less continent-size zones of Africa (Af), Australia and New Guinea (A&NG), Eurasia (EuAs), North American (NoAm), South America (SoAm) and South East Asia and Oceania (SEA&O). Their method requires that the relative amount of genera, and thus languages, in the sample for a macroarea is comparable to the relative amount of genera that the macroarea accounts for in the entire world, as in Table 1, where the numbers in the bottom two rows represent our present sample.

		Af	**A&NG**	**EuAs**	**NoAm**	**SEA&O**	**SoAm**	**Total**
world	# genera	74	140	43	92	66	106	521
	% genera	14.20	26.87	8.25	17.66	12.67	20.35	100.00
sample	# languages	23	43	13	28	20	33	160
	% languages	14.38	26.88	8.13	17.50	12.50	20.63	100.00

Table 1: Genus-macroarea sampling with a predetermined sample size of 160 (cf. Miestamo et al. 2016: 259)

[3] WBA; Isolate, South America.

We can use Eurasia to illustrate this principle. In Dryer (2013), this macroarea represents 8.25% of the world's genera (43/521). Accordingly, our 160-language sample should contain thirteen Eurasian languages – each from a different genus, of course – since the macroarea would then make up the similar proportion of 8.13% of the data (13/160).

Our sampling method takes two further steps, where possible, to reduce bias. First, when picking languages for a macroarea, priority is given to languages from genera that are not part of the same language family (Miestamo et al. 2016: 253). This step aims to ensure that smaller families, sometimes comprising only one genus, are represented – if the necessary information is available. Eurasia can again serve as an example: since our sample features Icelandic (ISL; Germanic, Indo-European), and we possess data for twelve entirely unrelated Eurasian languages, no other Indo-European genus/language is covered. However, it is not always feasible to eschew related languages. A selection of twenty languages from South East Asia and Oceania, for one, is highly likely to contain more than one Sino-Tibetan and Austronesian language, just because these language families account for the majority of genera in the macroarea. Second, the sampling method tries to avoid including any geographically adjacent languages (Miestamo 2016 et al. 2016: 249). To demonstrate this step, we can turn to Icelandic once more. One reason why this language is chosen to represent Germanic and not Swedish (SWE; Germanic, Indo-European) or Norwegian (NOR; Germanic, Indo-European) is that our sample also features Pite Saami, a Uralic language spoken in Sweden and Norway. It is not always desirable, though, to exclude neighboring languages altogether. For small regions with substantial linguistic diversity that forms a large proportion of a macroarea's genera (e.g. the Northern Territory in Australia), strict adherence to this second step would mean missing out on whole genera. We therefore go with Miestamo (2005: 32) in such situations and give precedence to genealogical rather than geographical variety.

The final prerequisite for a language to be part of the sample is particular to our study: it must possess both an imperative and a negative imperative. This requirement may seem trivial but Miestamo & van der Auwera (2007), for instance, consider North Slavey for their investigation into (a)symmetry in imperative negation. This language has a construction, in (6a) with the prohibitive marker *ʔehdḿí*, that is dedicated to expressing 'don't!'. In other words, there is a negative imperative in North Slavey. The primary way to get someone to do something in the language, however, is (6b). This construction is actually a declarative that is being used directively (cf. *you are*

going home! with a certain intonation in English) and North Slavey possesses no alternative that is more specialized to conveying 'go home!' or, put differently, no imperative. Relying on such a language for research into imperative negation does not seem felicitous: any (a)symmetries that would be established exist not between imperative and negative imperative but between negative imperative and positive declarative.

(6) North Slavey (SCS; Athapaskan, Na-Dene; Rice 1989: 1109)
   a. *ʔehdǵní ʔíyę hahʔá*
      PROH meat eat.2PL.IPFV
      'Don't y'all eat the meat!'
   b. *ʔáradíła*
      go.home.2SG.IPFV
      'Go home!' or 'You are going home.'

To exclude languages like North Slavey, one should ideally have clear cross-linguistic definitions/comparative concepts of the imperative and the negative imperative. As Jary & Kissine's (2016) in-depth discussion about imperatives shows, though, developing such definitions is far from straightforward. Going into the pros and cons of any proposal would take a considerable amount of space – which the present article, unfortunately, does not have (but see Van Olmen 2024: 212–220). The following characterization and examples will have to suffice here. For us, the (negative) imperative is a distinct grammatical construction, in morphological terms (see Tucano) and/or syntactic ones (see English *eat!*), that has no other prototypical function than to express an attempt by the speaker to get their addressee(s) (not) to do something (see also van der Auwera 2005: 565; Aikhenvald 2010: 1–2; Jary & Kissine 2016: 132). Consider now Ghomara in (7) and Lokono in (8). In the first language, there exists a construction that is dedicated to conveying directivity. This imperative in (7a) is marked by the lack of any inflection in the singular and the suffix *-w* in the plural. However, Ghomara's most basic strategy to issue a negative directive does not count as a negative imperative. The construction in (7b) has another typical function, i.e. the expression of the future declarative. The second language possesses neither an imperative nor a negative imperative. The constructions in (8a) and (8b) may be the primary ways in Lokono for a speaker to get an addressee (not) to do something but, like (6b), they tend to serve as present declaratives too.

(7) Ghomara (GHO; Berber, Afro-Asiatic; Mourigh 2015: 148)

a. *hala-ø(/w)*
come-2SG.IMP/2PL.IMP
'(Y'all) come!'

b. *ma ya kerz-et ši*
NEG IRR plough.AOR-2SG NEG
'Don't plough!' or 'You will not plough.'

(8) Lokono (ARW; Antillean Arawakan, Arawakan; Patte 2008: 105, 145)

a. *bu-shika da-mun no*
2-give 1-DAT 3.F
'Give it to me!' or 'You give it to me.'

b. *ma-iya-n b-a*
NEG/PRIV-cry-NMLZ 2SG-AUX
'Don't cry!' or 'You don't cry.'

Languages such as North Slavey, Ghomara and Lokono should, in our view, not be part of any study of (asymmetries in) imperative negation and they are indeed skipped in the compilation of the present article's sample.

For an overview of our sample, we refer to the Appendix 1. It provides, for each language, the following information: its macroarea, the language family that it belongs to, its genus and its Glottolog and ISO 639-3 codes.

2.2. Tense

As Aikhenvald (2014: 206) points out, "the most frequently attested grammaticalized time reference in imperatives is that of immediate versus delayed" compliance. Of the 160 languages in our sample, eighteen or 11.25% are found to make this type of distinction in their imperatives. In ten of them, it is expressed by the addition of a marker, like *-ri* in (9a),[4] and, in another five, by imperative markers that are in complementary distribution to each other, like *-git* and *-na* in (9b). West Greenlandic is

[4] The question whether such markers/distinctions relate to any declarative ones (of futurity) is the subject of Aikhenvald's (2014: 207–211) investigation and is of no concern to us here, as it has to do with (a)symmetry between imperatives and declaratives.

the sole language in our data with both strategies. Its complementary imperative suffixes occur only in the intransitive second person singular, however.

(9) West Greenlandic (KAL; Eskimo, Eskimo-Aleut; Fortescue 1984: 25–26)
   a. *uja(-ri)-sigik*
      look.for-DEL-2PL > 3PL.IMP
      'Y'all look for them (later)!'
   b. *ingin-niear-git(/na)*
      sit.down-CON-2SG.IMM.IMP/2SG.DEL.IMP
      'Sit down (later)!'

The way that Menggwa manifests the distinction is by means of different stems (for those verbs allowing the alternation, that is). As (10) shows, its imperative is characterized by the absence of tense-aspect-mood inflection and *sama* 'cook' is replaced by *dama* to express delayed compliance.

(10) Menggwa (KBV; Senagi; de Sousa 2006: 382)
   sama(/dama)-wa-a-ø
   cook/cook.FUT-2SG-3SG.F-IMP
   'Cook it (later)!'

Chinantec Lealao in (11), lastly, is somewhat unique in our sample. Not only does the language distinguish immediate from temporally vague (including delayed) compliance, it also draws on two completely different constructions to make the distinction.

(11) Chinantec Lealao (CLE; Chinantecan, Oto-Manuean; Rupp 1989: 93)
   a. *ŋia*^{M} *la*^{M}
      come.2SG.COMPL here
      'Come here (now!).'
   b. *ʔi*^{M} *ha*^{LM}*i*
      REL come.2SG.PROG
      'Come (sometime)!'

Both (11a) and (11b) are grammatically distinct: the first one's completive verb form does not ordinarily appear without further inflection and the second one's relative

marker *ʔiM* requires an antecedent normally. They are also both dedicated to conveying directivity. The difference between the two imperatives is that (11a) presumes a direct response and (11b) does not.

Eight of these languages retain the tense distinction in their negative imperative and, like Kunuz Nubian in (12), they all do so with the same marking as in their imperative – except for Edolo in (13), which has prohibitive counterparts to its immediate and delayed imperative suffixes.

(12) Kunuz Nubian (KZH; Nubian, Eastern Sudanic; Abdel-Hafiz 1988: 161–163)
   a. *ju(:-ka)-ø*
      go-DEL-2SG.IMP
      'Go (later)!'
   b. *jom(-kam)-me-ø*
      hit-DEL-NEG-2SG.IMP
      '(At a later point,) don't hit!'

(13) Edolo (ETR; Bosavi; Gossner 1994: 49)
   a. *molö gobe-mo(/malo)*
      food cook-IMM.IMP/DEL.IMP
      'Cook food (later)!'
   b. *ama-mabu(/mabio)*
      do-IMM.PROH/DEL.PROH
      '(At a later point,) don't do that!'

Koasati also preserves its positive contrast between immediate and delayed compliance in the negative, as (14a) and (14b) show. The language makes a rare additional distinction, however, with its further delayed imperative in (14c) and this construction has no negative equivalent.

(14) Koasati (CKU; Muskogean; Kimball 1991: 270–271)
   a. *ip-ø(-aḥ)*
      eat-2SG.IMP-DEL
      'Eat it (later)!
   b. *is-p-án(-naḥ)*
      2SS-eat-PROH-DEL
      '(At a later point,) don't eat it!'

c. *ip-ø-ạ́:hah*
eat-2SG.IMP-FUR.DEL
'Eat it much later!'

Complete neutralization occurs in seven of the eighteen languages. Most of them are like West Greenlandic in (15) and (9) in that their negative imperative does not resemble their imperative at all constructionally. The West Greenlandic one employs the negative contemporative forms of the verb, which are normally found in dependent clauses and whose independent use is dedicated to expressing 'don't!'. In just two languages do we see neutralization in a negative imperative that is similar to the imperative. Kolyma Yukaghir in (16) is one of them.

(15) West Greenlandic (Fortescue 1984: 27)
patin-nanga
hit-2SG > 1SG.NEG.CONTEMP
'Don't hit me!'

(16) Kolyma Yukaghir (YUX; Yukaghir; Maslova 2003: 140)
a. *jaqa-ŋi(-ge)-k*
arrive-PL-DEL-IMP
'Y'all arrive (later)!'
b. *el-l'aqa-ŋi(*-ge)-le-k*
NEG-arrive-PL-DEL-PROH-IMP
'Don't y'all arrive!'

For two more languages, finally, the available material does not allow us to determine whether the distinction between immediate and delayed compliance that exists in the positive is possible in the negative.

On the whole, roughly half of the languages in our sample with a tense distinction in the imperative neutralize it in the negative imperative. Moreover, no language seems to distinguish immediate from delayed compliance solely in its negative imperative. These observations suggest that there is a systematic asymmetry of neutralization from positive to negative here. One potential counterexample comes from a language that is not part of the present sample, Nyankore in (17) (see Van Olmen et al. 2023: 201–202).

(17) Nyankore (nyn; Bantu, Niger-Congo; Morris & Kirwan 1972: 10)

a. *o-ta(-ri)-gyend-a*
2SG-NEG-REM.FUT-go-FV
'(At a much later point), don't go!'

b. *mu-rya-gyend-a*
2PL-REM.FUT-go-FV
'Y'all go much later!' or 'Y'all will go much later.'

Its negative imperative in (17a) can convey delayed compliance by inserting the remote future marker. There does exist a positive equivalent to the construction with *-ri* but, as (17b) makes clear, it "is the same in form as the indicative far future" (Morris & Kirwan 1972: 10, who also point out that the negative imperative differs from its indicative counterpart in the position of the negative prefix). One may therefore argue that it does not constitute a "proper" imperative (see Section 2.1). It seems sensible, though, not to attach too much importance to the situation in Nyankore, since its interpretation depends heavily on what one takes (negative) imperatives to be.

2.3. Direction and/or location

As Aikhenvald (2010: 133–138) shows, imperatives frequently make space-related distinctions, often but not always as the only clause type in a language. They may indicate that the addressee is expected to move toward or away from the speaker to do something. These directions can be called andative and venitive respectively and are illustrated in (18). Imperatives may also signal that the addressee is supposed to do something close to or far from the speaker or simply at a different place. An example of such a location-specifying construction is Trio's so-called "dislocative" imperative with *-ta* in (19a). It tries to get the addressee to carry out the action elsewhere and is in complementary distribution with the ordinary and venitive imperative suffixes *-kë* and *-mïï* in (19b) (Carlin 2004: 307 explicitly writes that the latter is not a purely proximal imperative).

(18) Ese Ejja (ese; Tacanan; Vuillermet 2012: 666)
ixya(-ki/wa)-kwe
eat-AND/VEN-IMP
'(Go/come to) eat!'

(19) Trio (tri; Cariban; Carlin 2004: 307, 313)

a. *ene-ta*
look-DISLOC.IMP
'Look somewhere else!'

b. *ene-kë(/miï)*
look-IMP/VEN.IMP
'(Come) look!'

It is important to add here, with Aikhenvald (2014: 211–212), that tense distinctions in imperatives may acquire locational/directional connotations. A delayed imperative, for instance, can imply distance too. In some languages, the marking is even entirely vague between a temporal and a spatial interpretation. The Arawá suffix *-jahi* in (20) is a case in point and would have to be considered in this section as well as in Section 2.2. However, the language is not part of the present sample, which contains no similar cases.

(20) Arawá (aru; Arauan; Aikhenvald 2014: 211)

otara *noki* *ti-jahi*
1EXCL.OBJ wait 2SG-DEL/DIST.IMP.F
'Wait for us (in some distant time or place)!

Of the languages in our data, thirteen or 8.13% feature space-related distinctions like the above in the imperative. Most resemble Ese Ejja in that there is extra marking in the regular construction, like *-ki* and *-wa* combining with the imperative suffix *-kwe* in (18), to add a direction or a location to the directive. In the other six languages, we find marking that replaces the ordinary exponent of the imperative, as in Trio in (19), but half of them still possess the Ese Ejja strategy too. Nuuchahnulth in (21) can serve as an example.

(21) Nuuchahnulth (nuk; Southern Wakashan, Wakashan; Davidson 2002: 271, 296–297)

a. *haťi˙s=csu:*
bathe=2PL.AND.IMP
'Y'all go and bathe!'

b. *hič-ma-(č)i:ɬ-šiƛ = ʾiˑč(-ak)*
illuminate-thing-make-PFV = 2PL.IMP-VEN
'Y'all (come and) make torches!'

The language substitutes andative imperative clitics, *=csu:* in (21a), for the regular ones, *=ʾiˑč* in (21b), to express 'go and ...!'. The venitive meaning 'come and ...!', by contrast, is marked by simply attaching the suffix *-(a)k* to the normal imperative clitics, as in (21b).

Let us now turn to the negative imperative. We have only indirect evidence, in the form of an example, for just one of the thirteen languages above of a space-related distinction made in the positive also appearing in the negative: the Ese Ejja andative in (18a) and (22).

(22) Ese Ejja (Vuillermet 2012: 470)
a'a *akwi-kwi-jeyo = jo* *sowa-ki-xi*
PROH tree-plant-slippery = LOC go.up-AND-PROH
'Don't go up on this slippery plant!'

Similarly, for no more than two of these languages do we know, beyond reasonable doubt, that the negative imperative neutralizes the choices present in its positive equivalent. Djingili is one of them. Pensalfini (2003: 232) explicitly states that the only acceptable (negative) imperative forms in the language are those in (23): the regular imperative in (23a) (the absence of subject marking makes this irrealis construction dedicated); the andative one in (23b); and the negative one in (23c). In other words, the option in the positive of indicating a direction does not appear to exist in the negative.

(23) Djingili (jig; Djingili, Mirndi; Pensalfini 2003: 232)
a. *ngaja-mi*
look-IRR
'Look!'
b. *ngiji-yirri*
look-AND.IMP
'Go and look!'

c. *ngiji-ji*
look-PROH
'Don't look!'

The descriptions of the ten remaining languages do not address or are insufficiently clear about the question whether the space-related distinctions in the imperative are possible in the negative imperative. Carlin (2004: 309–311), for instance, writes that, in Trio, negative imperatives consist of a negated non-finite form of the lexical verb and the imperative of 'be', like in (24). For 'be', she explicitly mentions the regular imperative suffix *-kë* in (19b) but does not specify that the dislocative and venitive imperative endings in (19) are ungrammatical. One could interpret this information as pointing to neutralization (cf. Aikhenvald 2010: 184) but the evidence is far from conclusive.

(24) Trio (Carlin 2004: 309)
in-ene-ø-wa *eh-kë*
3-see-NFIN-NEG be-IMP
'Don't look at it!'

It is nevertheless worthy of note that, for so many languages, directional and/or locational differentiation is discussed only for the imperative and, furthermore, that no language in our sample appears to make such distinctions just in the negative imperative. It is also interesting that there is a common cross-linguistic path of change from 'go', whose meaning then bleaches, to imperative marking (see Mauri & Sansò 2011: 3497–3500) but that, to our knowledge (e.g. Aikhenvald 2010: 351–362), no path from '(not) go' to negative imperative marking has been established. Together, these observations can, in our view, still be argued to be indicative of an asymmetry of neutralization of space-related distinctions from positive to negative, as postulated by Aikhenvald (2010: 183–184) too.

2.4. Intersubjectivity

It should come as no surprise that, as inherently addressee-oriented constructions, imperatives in the world's languages often exhibit formal variation that one could characterize as intersubjective in nature. Intersubjective meaning is understood here as the

"explicit expression of the SP[eaker]/W[riter]'s attention to the 'self' of addressee/reader in both an epistemic sense (paying attention to their presumed attitudes to the content of what is said)" and, more importantly for us, "a more social sense (paying attention to their 'face' or 'image needs' associated with social stance and identity)" (Traugott 2003: 128). It manifests itself in the imperative as distinctions marking the interpersonal relationship between speaker and addressee (Aikhenvald 2010: 212–223) and/or the directive's illocutionary strength (Aikhenvald 2010: 203–212).

An example of an interpersonal distinction can be found in Kurtöp. The imperative suffix *-le* in (25a) is described as informal. It is employed between friends and people of similar social status or to issue directives to children. The so-called polite imperative ending *-lu* in (25b), by contrast, is used when the addressee has higher status or the speaker just wants to evoke a sense of respect.

(25) Kurtöp (xkz; Bodic, Sino-Tibetan; Hyslop 2011: 571, 568)
a. *gi-lu*
go-INFML.IMP
'Go!'
b. *dot-le*
sleep-POL.IMP
'Sleep!'

An example of a distinction in illocutionary strength comes from Kwazá in (26).

(26) Kwazá (xwa; Isolate, South America; Van der Voort 2004: 305)
koreja'ro *wa'ja-nỹ(-ca)-'ra*
pan bring-REFL-EMP-IMP
'(I'm telling you,) bring here the pan (I've asked you before)!'

The imperative in this language is indicated by the suffix *-'ra*. The marker *-ca* can be inserted before this ending and it has the effect of rendering the directive more emphatic or forceful, as the translation inside the parentheses in (26) aims to suggest.

Two comments are in order. First, languages do not always use dedicated markers, such as those in (25) and (26), to make intersubjective distinctions in the imperative. They also often co-opt other grammatical categories to express them (Aikhenvald 2010: 219–223). In Tukang Besi, for instance, the imperative differs from other clause types

in its lack of a subject prefix. A bare case like (27a) is perceived as slightly brusque, though. One way to soften the directive is to attach the perfective aspect suffix *-mo* with an exaggerated fall in pitch at the end, like in (27b). In the same vein, delayed imperatives are sometimes repurposed to convey less forceful and/or more polite directives, compliance with which need no longer be situated in the future (see Aikhenvald 2014: 210–211 too). Take Nungon, for instance: in this language, "the Delayed Imperative is politer than the Immediate Imperative" (Sarvasy 2017: 235) and, as evinced by (28), where immediate compliance is clearly expected, such intersubjective considerations may be the only motivation for the use of the delayed imperative.

(27) Tukang Besi (khc; Celebic, Austronesian; Donohue 1999: 453, 525)
a. *koka*
peel
'Peel!'
b. *kede-mo*
sit-PFV
'Sit down!'

(28) Nungon (yuw; Finisterre-Huon, Trans-New Guinea; Sarvasy 2017: 236)

karup,	*yii*	*ma-irök*	*mama-na,*	*wo-rok*
quick	vine	cut-2SG.DEL.IMP	mom-1SG.POSS	that-SEMB

'Quick! Cut the vine, my mom, that's it.'

Second, intersubjective distinctions in imperatives are not always simply a matter of adding or replacing some marker. They may also be expressed by distinct constructions. In Shangaci, for example, both the verb form missing a subject prefix in (29a) and the independent main clause use of the subjunctive verb form in (29b) are specialized for conveying directivity and constitute imperatives. They fulfill a different intersubjective function, however: (29b) is regarded as more polite than (29a) (see also Van Olmen et al. 2023: 206–210).

(29) Shangaci (nte; Bantu, Niger-Congo; Devos & Van Olmen 2013: 10, 15)
a. *khol-á*
grasp-FV
'Grasp!'

b. *u-khól-e*
2SG-grasp-SBJV
'Grasp please!'

What is crucial here is that these phenomena in Tukang Besi, Nungon and Shangaci are, in our view, as central to intersubjectivity in the imperative as the forms and variation found in Kurtöp and Kwazá. Accordingly, the present section will take all such patterns into account to see how (a)symmetric imperative negation is when it comes to intersubjective distinctions.

In our data, we have evidence for thirty-eight languages of imperatives marking such distinctions. They total 23.75% of our sample, a comparatively high percentage (cf. Sections 2.2 and 2.3) that could be seen as indicative of how central intersubjective concerns are to the imperative. Of these languages, twenty-two resemble Kwazá in (26) in that distinctions are made by adding markers, eleven are like Kurtöp in (25) in using markers that are in complementary distribution with one another and six are similar to Shangaci in (29) in employing different constructions. Looking at their imperatives' negative equivalents, we can observe that fifteen of the languages preserve the intersubjective distinctions in imperative negation. Perhaps unsurprisingly, in all but four of them, the negative imperative is constructionally symmetric vis-à-vis the imperative. Kurtöp in (30), with the negative prefix *ma-*, is a case in point.

(30) Kurtöp (Hyslop 2011: 318, 565)
a. *ma-lang-u*
NEG-be.full-INFML.IMP
'Don't be full of …!'
b. *ma-chak-e*
NEG-step-POL.IMP
'Don't step!'

An example of a language where there is no such symmetry but intersubjective distinctions are still maintained is Kayardild. In its imperative, the verb is marked in the same way as the "positive actual" but subject pronouns are optional in the construction and its case marking of objects is highly idiosyncratic (Evans 1995: 256), as the nominative third person singular in (31a) suggests. In its negative imperative in (31b),

the verb carries the prohibitive suffix *-n(a)* instead of "imperative" *-ja*. Crucially, *barri* 'just' can be appended to both constructions to soften the directive, as (31) shows, and this particle is, in fact, only found in (negative) imperatives.

(31) Kayardild (GYD; Tangkic; Evans 1995: 384)
  a. *barri* *wuu-ja* *ni-y*
     just give-IMP 3SG-NOM
     'Just give it back to him!'
  b. *barri* *kuliya-kuliya-n*
     just fill-REDUP-PROH
     'Just don't give me too much food!'

In fourteen other languages, however, the intersubjective distinctions made in the positive are neutralized in the negative. Perhaps not unexpectedly, nine of the languages have a negative imperative that is constructionally asymmetric vis-à-vis its positive counterpart. In Aguaruna, for instance, the regular imperative is marked by *-ta*, as in (32a), and the familiar imperative, which tends to be used with relatives and children, by singular *-kia* or plural *-khua*, as in (32b). None of these suffixes occurs in the negative imperative, which shares the ending *-i* with the apprehensive but differs from it in featuring the extra second person marker *-pa*, as in (32c). The construction makes no familiarity-based distinction.

(32) Aguaruna (AGR; Jivaroan; Overall 2017: 70, 72, 75)
  a. *su-sa-ta-hum*
     give-PFV-IMP-2PL
     'Y'all give!'
  b. *yu-wa-khua*
     eat-PFV-2PL.FAM.IMP
     'Y'all eat!'
  c. *ihu-i-pa-hum*
     stab-APPR-2-2PL
     'Don't y'all stab!'

In the five languages with constructional symmetry, neutralization may be a matter of the negative imperative simply not tolerating an intersubjective element that can

appear in the imperative (e.g. Telban 2017: 275 on Karawari's[5] intensifying marker *karka*). It may also concern the lack of a negative equivalent to one of the positive constructions. For example, of the options in (29), Shangaci can only negate the one deemed more polite, like in (33), but, in the negative, this subjunctive construction has no particular intersubjective associations anymore. Haida is another case in point. This language possesses an imperative marked by the clitic *=hl@* on the clause's first constituent and a familiar imperative marked by the affix *-.alaa*, as shown in (34a) and (34b) respectively. The former has a negative counterpart, like in (34b), but the latter does not.

(33) Shangaci (Devos & Van Olmen 2013: 24)
u-si-khol-e
2SG-NEG-grasp-SBJV
'Don't grasp!'

(34) Haida (HAI; Isolate, North America; Enrico 2003: 121, 126)
a. *daa=hl@* *gyaaxa*
2SG=IMP stand
'You stand up!'
b. *ga* *taa-.alaa* *gwáa*
INDF eat-FAM.IMP Q
'Eat, hey?'
c. *sgawsid-aay=hl@* *gam* *kidahl-rang*
potato-DEF=IMP NEG mash-NEG
'Don't mash the potatoes!'

Besides the twenty-nine languages discussed so far, we have nine for which intersubjective distinctions are mentioned just for the imperative. Four of them possess a constructionally asymmetric negative imperative, five a constructionally symmetric one. The descriptions, however, do not contain any information about or any examples of the positive distinctions being made in the negative. Consider Tukang Besi in (27) and (35) and Sandawe in (36).

[5] tzx; Lower Sepik, Lower Sepik-Ramu.

(35) Tukang Besi (Donohue 1999: 454)
bar(a) *(')u-kede* *i* *atu*
PROH 2SG.REAL-sit OBL there
'Don't sit there!'

(36) Sandawe (sad; Isolate, Africa; Steeman 2011: 105, 173, 259)
a. *pèé-é=kò*
put.SG-3=2SG.IMP
'Put it down!'
b. *í=↓kwáá*
come.SG=2SG.IMP
'Please come!'
c. *mèé=kò* *bô*
PROH=2SG.IMP say
'Don't say …!'

We do not know whether Tukang Besi *-mo* in (27b) can be attached to (35) too or whether, like the enclitic in (36a), Sandawe's "less imperative" alternative in (36b) can occur in the negative imperative in (36c) (Steeman 2011: 105).

In short, there is evidence for a tendency to neutralize intersubjective distinctions in (negative) imperatives and, in line with what is known from standard negation, it seems to go from positive to negative. Yet, our sample also includes four languages where such distinctions are made only in the negative (see Aikhenvald 2010: 189–190 too). Páez is one of them. The constructionally asymmetric negative imperative with *-nu* in (37a) has an equally asymmetric but less usual and more emphatic substitute marked by *-puʔn*, like in (37b). These options do not exist in the language's imperative in (37c).

(37) Páez (pbb; Isolate, South America; Jung 2008: 87–88)
a. *uʔx-nu-we*
go-PROH-2PL
'Don't y'all go!'
b. *vit-puʔn-we*
lose-EMP.PROH-2PL
'Don't y'all lose (it)!'

c. *m-dex-we*
IMP-sleep-2PL
'Y'all sleep!'

This type of neutralization occurs in 21.05% of the languages in our data with intersubjective distinctions in the negative imperative (i.e. four like Páez versus fifteen like Kurtöp). Neutralization in the other direction *is* much more frequent, though – arising in 48.28% of the sample languages with intersubjective distinctions in the imperative (i.e. fourteen like Aguaruna versus fifteen like Kurtöp). For that reason, although there is clearly no unidirectional asymmetry of neutralization in the intersubjective domain, we can still conclude that, cross-linguistically, this type of asymmetry is more likely from positive to negative than vice versa.[6]

2.5. Interim summary

The findings of this section's typological survey confirm that tense in imperative negation exhibits a systematic asymmetry of neutralization from positive to negative. Distinctions in the imperative to do with the time of compliance may and often do indeed disappear in the negative imperative but the reverse does not seem to happen. Our results are highly suggestive too of a similar asymmetry in the marking of direction and/or location in imperative negation. Distinctions concerning the addressee's movement or the place of compliance are typically mentioned only for the imperative and never just for the negative imperative. For a couple of languages at least, we also have clear indications of actual neutralization from positive to negative. For intersubjectivity in imperative negation, lastly, the results are more ambiguous.[7] As already

[6] One reviewer rightly indicates that the difference between neutralization from positive to negative and neutralization from negative to positive is not statistically significant. However, the result of their Fisher's exact test, i.e. $p = 0.073$, can still be interpreted as a trend, which may be seen as receiving some further support from the fact that there are an additional nine languages for which intersubjective distinctions are mentioned for the imperative but simply not discussed for its negative counterpart.

[7] One of the reviewers wonders whether "one reason" is "that negation itself is intersubjective in nature". We do not at present have an obvious answer to this interesting question (or, for that matter, a clear explanation for the findings on intersubjectivity in general, as discussed in Section 3.4) but do wish to mention that, to us, any intersubjectivity of negation would seem quite different from the types of distinctions of concern here: 'don't!' does not directly mark either interpersonal relations or illocutionary strength.

shown by Aikhenvald (2010: 189–190) with Manambu (mle; Ndu, Sepik), negative imperatives can make more intersubjective distinctions than imperatives. Our numbers still suggest, however, that such asymmetry does not occur as frequently in the world's languages as its opposite.

3. Usage-based perspective

This section will first discuss our corpus material (Section 3.1). Next, we will examine whether the asymmetries in tense (Section 3.2), direction and/or location (Section 3.3) and intersubjectivity (Section 3.4) have any basis in usage. An interim summary will be given at the end (Section 3.5).

3.1. Corpus data

For our usage-based perspective, the focus will be on two languages, i.e. English and Dutch. While we acknowledge that this choice has its limitations, in that the languages are very closely related and their cultures are probably quite similar too, our motivation for it is two-fold. First, a study examining the ways that (negative) imperatives are employed in discourse requires extensive familiarity with the languages under investigation, which we have for English and Dutch (e.g. Van Olmen 2011, 2019). Second, research exploring whether cross-linguistic grammatical differences between imperatives and negative imperatives have a basis in usage should ideally look at languages where those differences are not part of the grammar. English and Dutch fit this description, for the most part. Neither language makes grammatical distinctions in its (negative) imperative between immediate and delayed compliance or relating to the location of compliance. The expression of the addressee's movement, by contrast, does seem to have grammaticalized to some extent. Nicolle (2009: 187–189, 196–200) shows for English that *go/come*-V(erb) in (38a) is a different construction than *go/come-and*-V in (38b) (e.g. *she went and visited him* versus **she went visited him*). He also argues that "*go*-V developed diachronically from *go-and*-V in the context of imperative clauses (like 38c), whilst *come*-V may have developed either by analogy with *go*-V or as a result of an independent development from *come-and*-V" (Nicolle 2009: 204) and that *go/come*-V has undergone subjectification – in the sense of Langacker (1990) – as "the subjective component of meaning [i.e. the

speaker as the deictic center of the movement] … is incorporated into the representation of the whole event" (Nicolle 2009: 203–204).

(38) a. She will go/come visit him.
b. She will go/come and visit him.
c. Go (and) see it!

The question crucial for our purposes, though, is whether *go/come*-V, as well as *go/come-and*-V and other similar constructions, is restricted to imperatives or, put differently, whether there is a grammatical asymmetry here. The corpus examples of negative imperatives in (39) suggest that the answer is no.

(39) a. Don't go see this movie based on the fact it's labeled a thriller.
(enTenTen20: 2593103)
b. Don't come read with me. I am mad at you, and I will tuck my own self in.
(enTenTen20: 44173818)
c. Don't go and glean in another field and don't go away from here.
(enTenTen20: 22372049)
d. A fantastic pub right in the heart of soho. Don't come and ruin it.
(enTenTen20: 8371733)

In the same vein, English and Dutch (negative) imperatives do not exhibit any conventionalized differences in intersubjective marking either, to our knowledge. The linguistic elements known to be able to modify illocutionary force and/or mark interpersonal relationships – such as *please* and tag questions in English (e.g. Wichmann 2004; Kimps & Davidse 2008) and modal particles and the formal second person imperative subject *u* in Dutch (e.g. Vismans 1994; Fortuin 2004) – can all appear in both the imperative and the negative imperative. Probably the only obvious exception is *do*-support in English. It is an option in imperatives and tends to emphasize whatever function they are fulfilling (cf. the offer *do have a cookie!* and the order *do shut up!*; De Clerck 2006: 330–332) but, in negative imperatives, *do* is simply required by *not* and it does not contribute anything to their meaning. It is important to bear in mind, however, that *do*-support is very infrequent in the imperative (see De Clerck 2006: 172, who detects it in just 1.90% of his 1,580 corpus attestations) and its impact on any usage data will therefore be limited.

Our data comes from two main sources. The first one is Van Olmen's (2011) earlier study of the illocutionary functions of (negative) imperatives in comparable corpora, one of speech and one of plays. The former consists of the spoken part of the International Corpus of English Great Britain (ICE-GB; Survey of English Usage 2006) – ca. 600,000 words of different types of private and public dialogue and scripted and unscripted monologue from the 1990s – and a selection of the Northern Dutch files of the Corpus Gesproken Nederlands (CGN; Nederlandse Taalunie 2004) that closely mirrors the composition of the ICE-GB – ca. 300,000 words from the late 1990s and early 2000s (see Van Olmen 2011: 55–56, 59–61). The latter is made up of plays all written by different speakers of British English and Northern Dutch and all translated by different speakers of Northern Dutch and British English respectively. This last feature was essential for Van Olmen (2011), who also exploited the plays as a parallel corpus, but restricted the number of works to choose from considerably. As an inevitable result, only one of the ten plays is authored by a woman and the corpus spans over 30 years, from 1974 to 2004 (see Van Olmen 2011: 115–117). These weaknesses notwithstanding, we can and will still use the source texts (i.e. not the translations) – totaling ca. 96,000 words for English and 70,000 for Dutch – as a comparable corpus here, inter alia, because they feature a comparatively high amount of (negative) imperatives, as Table 2 makes clear.

	Speech		**Plays**		**Total**	
	English	**Dutch**	**English**	**Dutch**	**English**	**Dutch**
Imperatives	738	250	596	288	1,334	538
Negative imperatives	119	15	131	74	250	89

Table 2: Absolute frequencies of the (negative) imperative in Van Olmen's (2011) corpus data.

These cases will constitute the core dataset of the present study. Note, though, that they do not include what Van Olmen (2011), following De Clerck (2006: 44– 45), calls "minor" (negative) imperatives. This group comprises instances that look like and originate from full-fledged (negative) imperatives but lack the ability to appear as autonomous, discursively prominent utterances and/or exhibit little formal and functional flexibility. Space does not allow an in-depth discussion of the distinction (see Van Olmen 2011: 34–36, 2019: 148–149). We hope therefore that the following list of examples will give the reader an adequate idea of the discourse markers, idiomatic phrases and such excluded from Table 2: English *come on!* 'oh no!', *don't mention*

it 'you're welcome' and *say* 'for instance', Dutch … *en noem maar op* '… and all the rest' (lit. '…and just name any!'), *kijk/zeg,* … 'look/say, …' and *pak hem beet* 'approximately' (lit. 'grab him!').

Our second source of data is the TenTen corpus family (Jakubícek et al. 2013) and will be used mainly for automated searches. It contains large bodies of texts, with billions of words, that "can be regarded as comparable corpora" as the same "technology specialized in collecting only linguistically valuable web content" is applied to build a corpus for each language in the family.[8] TenTen's diversity of discourse types (e.g. not only Wikipedia pages and newspaper articles but also online fiction and discussion forums) and sheer magnitude guarantee a certain degree of representativeness and a substantial number of hits for any queries. The corpora also have the benefit of being tagged with parts of speech, which makes it much easier to look for constructions like the (negative) imperative. Relying on web-crawled data comes with drawbacks too, of course. It is, for instance, hard to control for language variety (e.g. British/American English, (non-)native Dutch) or time. Still, to ensure at least some level of comparability with Van Olmen's (2011) corpora, we will restrict our searches of the enTenTen20 and nlTenTen20 data (both collected in 2020) to, respectively, .uk domains (2,899,739,619 words) and .nl domains (4,439,356,346 words).

Before looking at the corpus data in detail, let us draw attention to an interesting difference between the imperative and its negative counterpart in Table 2: in both English and Dutch, the negative imperative occurs much less often than its positive equivalent. Dutch speech displays the largest disparity, with approximately seventeen imperatives for each negative imperative, and the Dutch plays the smallest one, still with a ratio of almost four to one. If this difference in frequency is a trait of (negative) imperatives across the world's languages, it might partially explain the asymmetries of neutralization discussed in Section 2. As Miestamo (2005: 205–206) argues, "the lower frequency of marked categories (in this case negation) may have the effect of shaving off distinctions or preventing them to arise in the first place" since "it is not as economic to maintain a large number of distinctions in an infrequency category than it is in a more frequent one" (see also Haspelmath 2008, 2021). However, this potential impact of (in)frequency is difficult to prove and it remains fairly vague as a motivation. Moreover, one could also easily contend that economy can work against neutralization in particular for common distinctions in a language. If its negative imperative – unlike its imperative – did not allow them, it would actually be "an extra burden for language

[8] See https://www.sketchengine.eu/documentation/tenten-corpora/ (accessed 2023.04.28).

users to remember this special restriction with [this] … particular category" (Miestamo 2007: 308). It therefore seems sensible to consider (in)frequency as a possible contributing factor to our asymmetries rather than as *the* explanation for them.

3.2. Tense

As discussed in Section 1, Miestamo (2005: 211) attributes the frequent neutralization of tense-aspect-mood and person-number-gender distinctions in standard negation to discourse presuppositionality: as negative declaratives tend to be uttered in discourse environments where their positive equivalents are assumed or present in some way, the speech participants may be taken to be familiar with the 'when', 'who' and the like of their content already and there is less of a need to spell out those features. Intuitively, this explanation seems to be relevant for imperative negation as well: when one says 'don't X!' to someone, they are typically already Xing, in the context, or one has reason to think, based on the context, that they mean to X (see Miestamo & van der Auwera 2007: 71–72 too). In other words, discourse presuppositionality may also be a motivation for asymmetry in tense established in Section 2.2.

Importantly, discourse presuppositionality's role in negation is, in essence, a presumed discourse preference. With Miestamo et al. (2022: 135), we would therefore expect it to manifest itself in every language, at least in usage. A more specific hypothesis relating to tense in imperative negation, echoing Miestamo et al.'s (2024: 11–12) suggestion for standard negation, would then be that negative imperatives feature fewer temporal expressions than imperatives. To put it to the test, we can count how many of the (negative) imperatives in Table 2 contain lexical items or longer structures indicating a time of compliance in one way or another.[9] Table 3

[9] One of the reviewers finds this characterization of the expressions in question "rather imprecise" and, relatedly, takes issue with *quickly* in (40c). We acknowledge that our definition is fairly loose but believe that *quickly* nicely illustrates why it is phrased in this way. In Hasselgård's (2010: 39) semantic classification of (English) adjuncts, this adverb probably belongs to the category of manner instead of that of time. It would therefore have to be ignored if we restricted ourselves to temporal adjuncts in the strict sense (as Miestamo et al. 2024: 39 appear to do). However, discounting *quickly* in (40c) does not seem felicitous to us. In this example, the adverb does not express that having a look should happen in a fast way (at any point in time). Rather, the speaker uses it to urge the addressee to have a look at the time of speaking. In other words, *quickly* constitutes an expression of immediate compliance here and should be taken into account in our view. Our loose definition allows for its inclusion, just like it allows for the inclusion of the majority of cases that would count as straightforward adjuncts of time and of temporal position in particular, like those in (40a), (40b) and (40d).

gives the results in absolute numbers and percentages and (40) offers some English and Dutch examples.

	Speech		Plays		Total	
	English	**Dutch**	**English**	**Dutch**	**English**	**Dutch**
Imperatives	50 / 738	9 / 250	9 / 596	8 / 288	59 / 1,334	17 / 538
	6.78%	3.60%	1.51%	2.78%	4.42%	3.16%
Negative imperatives	10 / 119	0 / 15	2 / 131	0 / 74	12 / 250	0 / 89
	8.40%	0.00%	1.53%	0.00%	4.80%	0.00%

Table 3: Temporal expressions in the (negative) imperative in Van Olmen's (2011) corpus data

The numbers in Table 3 are, all in all, relatively low. To access more data, we can consult enTenTen20 and nlTenTen20. Locating (negative) imperatives in these corpora is not straightforward, though. The reason is that the English and Dutch constructions possess no dedicated morphology and can essentially only be defined in syntactic terms that, even in part-of-speech-tagged data, are hard to operationalize (e.g. the typical absence of the subject; verb-first word order; see Van Olmen 2011: 17–31). More open-ended searches are therefore bound to produce (too) many irrelevant hits (to be reliable without manual checking). At the same time, to ensure that only actual (negative) imperatives are retrieved, one cannot but fall back on more specific queries that will inevitably exclude relevant instances too. In our view, this second approach is the more suitable one for our purposes. Our rationale is two-fold. First, it allows us to collect data in an automatic way. Second, if the query for negative imperatives incorporates the same constraints as that for imperatives and if those constraints do not affect the phenomenon under investigation (e.g. the occurrence of temporal expressions), we can still compare the two constructions.

For imperatives in enTenTen20, for instance, we started with the query in (41a). It looks for the "base" form of all verbs (e.g. *go* and not *goes*, *went* and the like) except for *let*, to avoid non-second-person constructions such as *let's go*. Note that it rules cases like *let me go* 'allow me to go' out as well, of course. In addition, the SENTence-break punctuation at the beginning limits the search to verb-first sentences and *please* immediately preceding the verb restricts the hits further to likely imperatives (although it obviously excludes uses of the construction that are incompatible with the adverb). Next, to remove any negative imperatives from the results for (41a), we filtered out the hits corresponding to (41b). This query mirrors the one for imperatives

(i.e. the initial punctuation, the presence of *please*, the base form of *do*) but adds *not* and it was also used afterward to search for negative imperatives in enTenTen20 separately. Crucially, to keep the results as similar as possible, we then did away with all hits for this separate query of (41b) that feature *let*: since (41a) does not look for cases like *let me go*, we should not include cases *don't let me go* either.

(41) a. [tag = "SENT.*"] [lemma = "please"] [tag = "VV.*|VB.*|VH.*" & lemma! = "let"]
b. [tag = "SENT.*"] [lemma = "please"] [tag = "VV.*" & lemma = "do"] [lemma = "not"]

These searches produced 237,651 results for the imperative and 11,643 for the negative imperative in the .uk domain of the corpus. As a final check of their validity, we looked at a random sample of one hundred hits for each dataset and they were all found to be, respectively, imperatives and negative imperatives.

For (negative) imperatives in nlTenTen20, numerous attempts and modifications aimed at reducing the number of irrelevant hits while maintaining a substantial recall resulted in the query in (42). It essentially looks for sentences that are no longer than eleven words, begin with a verb stem and finish with an exclamation mark. Cases where the second word was *ik* 'I' were filtered out and, for imperatives, so were cases containing *niet* 'not', *geen* 'no', *niemand* 'nobody', *niets/niks* 'nothing', *nooit* 'never' or *nergens* 'nowhere'. The latter were taken to be the negative imperatives.

(42) <s> [tag = "verbpressg.*" & lemma! = "laten|kunnen|mogen|moeten|zullen|danken" & word! = ".*t|.*T|ben|BEN|Ben|bEn|beN|BEn|bEN|BeN|is|IS|Is|iS"] [tag = "adj.*|adv.*|det.*|int.*|noun.*|num.*|partte.*|prep.*|pron.*"]{0,10} [word = "\!"] </s> within <s/>

The searches yielded 195,567 results for the imperative and 7,066 for the negative imperative in the .nl domain of the corpus. These hits still include some false positives, such as (43).

(43) a. *Klaar voor de star!*
'Ready for the star!'
(nlTenTen20: 9987488)

b. [*Ik*] *Heb er zooo geen zin in!*
'[I] Am sooo not in the mood for it.'
(nlTenTen20: 9610555)

Note, however, that, in a random sample of one hundred instances for each dataset, we only found three that did not constitute an imperative and two that were not negative imperatives.

To compare the occurrence of temporal expressions in these (negative) imperatives, we focused on a selection of items – i.e. English *later*, *immediately, soon*, *today*, *tomorrow*, *tonight* and *when* and Dutch *later* 'later', *onmiddellijk* 'immediately', *gauw* 'soon', *vandaag* 'today', *morgen* 'tomorrow', *overmorgen* 'the day after tomorrow' and *vannacht* 'tonight' – and filtered the hits that contain them.[10] For English, the search window was kept narrow, to minimize the risk of irrelevant hits: a maximum of two words after the string in (41a) (e.g. *please visit her today*) and three words after the string in (41b) (e.g. *please don't visit her today*). For Dutch, we looked between the initial stem and the final exclamation mark of the query in (42). Table 4 presents the results in absolute terms and proportions and (44) gives some examples.

	English	**Dutch**
Imperatives	2,422 / 237,651	3,403 / 195,567
	1.02%	1.74%
Negative imperatives	28 / 11,643	6 / 7,066
	0.24%	0.08%

Table 4: Temporal expressions in the (negative) imperative in enTenTen20 and nlTenTen20

[10] We agree with one of the reviewers that, ideally, this selection should have been based (at least partly) on frequency data on temporal adjuncts. This information does exist at a general level (e.g. Biber et al. 1999 and Hasselgård 2010 on English) but, to our knowledge, there is little data on adjuncts of time in the (negative) imperative specifically (the fact that they are very infrequent there, as Table 3 shows, may play a role). Therefore, the current selection – though in part inspired by the expressions attested in Van Olmen's (2011) corpus data – has to remain somewhat intuitive. Relatedly, certain readers may wonder why 'now' and 'then' are not included here. The reason is that they are highly multifunctional items in both English and Dutch (negative) imperatives and, instead of conveying a temporal meaning, it frequently has intersubjective effects (see also Miestamo et al. 2024: 16–17). Consider, for instance, affectionate *now* in *don't worry now* or reinforcing *nou* in the delayed imperative in (40d).

(44) a. Please apply **immediately** to be considered for the role.
(enTenTen20: 2515462)

b. *Probeer het **morgen** weer!*
'Try again tomorrow!'
(nlTenTen20: 561207)

c. Please don't beat me **when I get home**.
(enTenTen20: 28675637)

d. *Neem **vandaag** zeker geen GSM s mee!*
'Definitely don't take any cellphones with you today!'
(nlTenTen20: 5796445)

Relatively speaking, the numbers are again quite low, with percentages ranging from 0.08% to 1.74%. However, in both English and Dutch, the negative imperative is found to occur significantly less often with temporal expressions than the imperative (respectively, χ^2 (df 1) = 69.15 with $p < 0.00001$ and χ^2 (df 1) = 112.95 with $p < 0.00001$). This fact could be seen as a reflection in usage of negation's discourse presuppositionality and thus, indirectly, as an explanation for the asymmetry in tense established for imperative negation cross-linguistically in Section 2.2.

Let us nevertheless have a more in-depth look at immediate versus delayed compliance. In our view, the most suitable corpus for such an investigation is the English and Dutch plays: they offer the explicit context necessary to determine time of compliance, do not contain any unintelligible passages and, for Dutch in particular, have a reasonable number of negative imperatives. For each language, we thus analyzed all negative imperatives in the plays and a random sample of imperatives of the same size. Examples in which the (negative) imperative involves immediate and delayed compliance are given in (45) and (46) respectively.

(45) a. Annie: Touch me then. They'll come in or they won't. Take a chance. **Kiss me**.
Henry: For Christ's sake.
Annie: Quick one on the carpet then.
(English plays, Tom Stoppard's *The Real Thing*)

b. *Vader:* *(wil het geld van Jurgen afpakken)*
Jurgen: *(weert Vader af)* ***Raak me niet aan***-
Vader: *(duwt Jurgen achteruit)*
'Father: (wants to take the money from Jurgen)
Jurgen: (fends off Father) Don't touch me-
Father: (pushes Jurgen back)'
(Dutch plays, Jeroen van der Berg's *Blowing*)

(46) a. Olive: Are the – er – are the Emersons coming round?
Anthea: Ah. Thereby hangs a tale. Possibly. I've asked them.
Olive: Oh, are they …?
Anthea: Oh dear. Well, [...] If they do come, **don't whatever you do ask after Christopher**.
(English plays, Alan Ayckbourn's *Joking Apart*)

b. *Hannah:* *Hij ging op het bed zitten, het kistje tussen zijn benen, ik knielde voor hem op de grond …*
Athalie: *En?*
[…]
Theodor: *Laat haar met rust. Jij begrijpt ook niets van vrouwen. […]* ***Zeg tegen Sylvia dat ze die man er niet meer in laat.*** *Hij is gevaarlijk. Hoor je me?*
Athalie: *(die naar Hannah luisterde) Ja … ik luister.*
'Hannah: He sat on the bed, the little box between his legs, I knelt on the floor in front of him …
Athalie: And?
[…]
Theodor: Leave her alone. You don't understand women at all. […] Tell Sylvia [who is not present] not to let that man in again. He is dangerous. Do you hear me?
Athalie: (listening to Hannah) Yes … I'm listening.'
(Dutch plays, Lodewijk de Boer's *The Buddha of Ceylon*)

There are, however, also numerous cases where the time of compliance is vague. In (47a), for instance, David's request to act as usual around his mother relates not to

any specific moment but to any future interaction with her. The negative imperative in (47b) too pertains to a longer (technically infinite) stretch of time.

(47) a. Xenia: You will tell me what I can do? Nursing, washing, anything.
David: Thank you, but there is nothing.
Xenia: […] We mustn't stay here gossiping. She must have rest and quiet.
David: You've forgotten what else I said. **Please behave as you normally would**. Otherwise you'll frighten her and aggravate her condition.
(English plays, Edward Bond's *Summer*)

b. *Sjaak:* *Hij maakt zich hier totaal onmogelijk! Ik begrijp ook niet dat jij dat maar steeds weer goed praat. Je bent toch niet blind Rooie…?*
De Rooie: *Misschien verandert-ie nog wel…*
Sjaak: *Ik heb geen enkele hoop. Rooie,* ***laat die jongen nooit 'n aanleiding worden dat er tussen ons een breuk komt.***
'Sjaak: He is making himself completely unbearable here! I also do not understand why you are always making excuses for that. You are not blind, are you, Rooie…?
De Rooie: He might still change…
Sjaak: I have no hope. Rooie, never allow that boy to become the reason for a rift between us.'
(Dutch plays, Gerard Lemmens's *Souvenirs*)

The distribution of the types in (45) to (47), in absolute and proportional terms, is presented in Table 5, separately for English and Dutch and for the imperative and its negative counterpart.

	English			**Dutch**		
	Immediate	**Delayed**	**Vague**	**Immediate**	**Delayed**	**Vague**
Imperatives	90	15	26	41	21	12
	68.70%	11.45%	19.85%	55.41%	28.38%	16.22%
Negative imperatives	71	12	48	40	8	26
	54.20%	9.16%	36.64%	54.05%	10.81%	35.14%

Table 5: Compliance in the (negative) imperative in Van Olmen's (2011) corpus data for plays

In English as well as Dutch, the negative imperative differs significantly from its positive counterpart (χ^2 (df 2) = 9.11 with $p < 0.05$; χ^2 (df 2) = 10.00 with $p < 0.05$ respectively). What it has in common in particular in the two languages is a comparatively higher number of vague instances. In other words, the negative imperative appears to be used more often than the imperative for situations where the time of compliance is less specific (36.64% versus 19.85% in English, 35.14% versus 16.22% in Dutch). This phenomenon may be taken as an additional or alternative explanation to negation's discourse presuppositionality for the cross-linguistic tendency to neutralize tense distinctions in negative imperatives: immediate versus delayed compliance is simply less relevant for them. One can also make sense of this apparent property at a more general level. What a speaker essentially wants to accomplish with a negative imperative is a situation where their addressee is not doing something and the absence of an event is more likely to be a continuous or continuing state than the realization of an event (cf. Miestamo 2005: 195–196 on the stativity of standard negation). If your interlocutor expresses anxiety about something and you tell them not to worry about it, for example, your initial aim may be to reassure your addressee there and then but the state of non-worry that you wish to achieve in them is probably intended to extend into the foreseeable future.

3.3. Direction and/or location

Section 2.3 suggests that there exists a cross-linguistic asymmetry in the marking of direction and/or location between imperatives and negative imperatives. When an imperative makes such distinctions, its negative counterpart may make them too but does not typically seem to do so. Moreover, the opposite situation does not appear to occur at all. The question that we wish to answer here is whether this phenomenon reflects usage in English and Dutch.

Adopting the same approach as in Section 3.2, we count the number of (negative) imperatives in Table 2 containing expressions of a direction and/or location for the addressee's (non-)realization of the event. The results are given in Table 6 in absolute numbers and percentages and (48) offers some examples. It is probably important to add, though, that cases such as (49) are not included in our sums. This imperative may contain auxiliary *gaan* 'go' but, as it often does, the verb conveys transition ('up') rather than motion ('go and stand') here.

	Speech		Plays		Total	
	English	**Dutch**	**English**	**Dutch**	**English**	**Dutch**
Imperatives	22 / 738	9 / 250	14 / 596	12 / 288	36 / 1,334	21 / 538
	2.98%	3.60%	2.35%	4.17%	2.70%	3.90%
Negative imperatives	0 / 119	0 / 15	0 / 131	1 / 74	0 / 250	1 / 89
	0.00%	0.00%	0.00%	1.35%	0.00%	1.12%

Table 6: Directional and locational expressions in the (negative) imperative in Van Olmen's (2011) corpus data

(48) a. Well do it **somewhere else**.
(ICE-GB: S1A.010.154)

b. If you want to acquire stock, **go and** talk to her.
(English plays, Howard Brenton & David Hare's *Pravda*)

c. *Nee **kom** maar niet kijken.*
'No, just don't come and watch.'
(Dutch plays, Lodewijk de Boer's *The Buddha of Ceylon*)

(49) *Gaat u weer even staan, moeder.*
'Please stand up again for a moment, mother.'
(Dutch plays, Joop Admiraal's *You are my Mother*)

It is evident from the figures in Table 6 that the (negative) imperative rarely features directional or locational expressions in English or in Dutch. Given these frequencies, it is unsurprising that there also exist no statistically significant differences between the imperative and its negative counterpart, in either corpus or either language. The almost complete absence of such expressions in negative imperatives, compared to their occasional appearance in imperatives, is nevertheless striking and perhaps telling.

For the larger enTenTen20 corpus, we relied on the queries in (41) to extract (negative) imperatives and filtered the results first for those containing the string in (50) and then for those with the lemmas *here* and *there*. As in Section 3.2, both searches were limited to a window of two words after the hit for the imperative and three for the negative imperative.

(50) [lemma = "come|go"] [lemma = "and"]? [tag = "V.*"]

The former filter gives us an idea of the amount of (negative) imperatives conveying direction, the latter filter an idea of those expressing location in English. Table 7 presents the findings in absolute and proportional terms and some examples can be found in (51).

	Direction	**Location**
Imperatives	715 / 237,651	2,368 / 237,651
	0.30%	1.00%
Negative imperatives	9 / 11,643	40 / 11,643
	0.08%	0.34%

Table 7: Directional and locational expressions in the (negative) imperative in enTenTen20

(51) a. Please **go and** read it and then pop back here.
(enTenTen20: 116918513)
b. Please don't camp **here** as it rightly annoys the local inhabitants.
(enTenTen20: 80977002)

Like in Table 6, the numbers are very low, both for expressions of direction and for expressions of location. Still, the negative imperative has significantly fewer of them than its positive equivalent (χ^2 (df 1) = 19.16 with $p < 0.0001$ for direction; χ^2 (df 1) = 49.46 with $p < 0.00001$ for location), which suggests that the apparent differences in English speech and drama are probably not accidental either.

For nlTenTen20, we first looked at the (negative) imperatives from Table 4, searched for with the query in (42) and the additional steps described there, and filtered the results for those containing the locational lemmas *hier* 'here' and *daar* 'there'. This operation produced 17,420 hits for the imperative and 295 hits for the negative imperative. In many of them, however, *hier* and *daar* are part of a so-called pronominal adverb, standing in for a prepositional constituent, like *hier … aan* 'with this' in (52). We therefore checked all negative imperatives by hand and kept only the 85 instances where the adverbs actually convey location, as in (54a). We did the same for a random sample of 295 imperatives and extrapolated the 91.86% of relevant cases to the total number of hits, giving us the speculative number of 16,003. For direction, exploratory searches indicated that *gaan*'s potential aspectual meaning in (49) would make any comparison without an in-depth semantic analysis unreliable. We thus decided to focus on *komen* 'come' here. Moreover, as the query in (42) does not allow for the infinitives that follow this auxiliary, like *uitproberen* 'try out' in (54b), we ran the adjusted one in (53) (but

adopted the same procedure as before to separate positive and negative imperatives) and filtered the results for those featuring an initial *kom* and an infinitive somewhere in the hit. The findings are given in Table 8 in absolute numbers and in percentages.

(52) *Verspil je tijd **hier** niet **aan!***
'Don't waste your time with this!'
(nlTenTen20: 11855191)

(53) <s> [tag="verbpressg.*" & lemma!="laten|kunnen|mogen|moeten|zullen|danken" & word!=".*t|.*T|ben|BEN|Ben|bEn|beN|BEn|bEN|BeN|is|IS|Is|iS"] [tag="adj.*|adv.*|det.*|int.*|noun.*|num.*|partte.*|prep.*|pron.*|**verbinf.***"] {0,5} [word="\!"] </s> within <s/>

	Direction	**Location**
Imperatives	939 / 145,782	16,003 / 195,567
	0.64%	8.18%
Negative imperatives	7 / 6,004	85 / 7,066
	0.11%	1.20%

Table 8: Directional and locational expressions in the (negative) imperative in nlTenTen20

(54) a. *Graaf **hier** geen kuil!*
'Don't dig a hole here!'
(nlTenTen20: 10983847)
b. ***Kom** het maar eens uitproberen!*
'Just come and try it out!'
(nlTenTen20: 729259)

The directional and locational expressions' frequencies are again low but the imperative nonetheless possesses significantly more of them than its negative counterpart (χ^2 (df 1) = 25.91 with $p < 0.00001$ for direction; χ^2 (df 1) = 454.56 with $p < 0.00001$ for location). In other words, the scarcity of such expressions in the negative imperatives in Dutch speech and drama does not appear to be a coincidence.

In summary, the corpus data for English and Dutch suggests that there exists a discourse preference for less directional and/or locational marking in the negative imperative than in the imperative. The typological findings in Section 2.3, pointing to a tendency to neutralize such distinctions from positive to negative, can reasonably be argued to reflect this preference. One way to account for it, with Miestamo (2005), is negation's discourse presuppositionality. When you try and get someone not to do something, they are often doing it at the time or you believe that they are planning on doing it. In other words, the positive is somehow already present in the discourse and explicating all of its details, including its direction and location, is thus less necessary in the negative. This explanation is quite general, though, as it can be applied to any area of neutralization in negation. We would therefore like to add that the discourse preference at issue, as well as its associated cross-linguistic tendency, may also be motivated by the relative inconsequentiality of direction and location in negative directive speech acts. In our view, if you attempt to get someone to stop or refrain from doing something, it will typically be less important to you, or to them, *where* the action does not take place than the action simply not taking place. Admittedly, it is not impossible to think of situations where direction or location could be relevant in a negative directive. For instance, if a speaker wants their addressee to stay or move toward them and do something and if they really wish to exclude the alternative, they might conceivably say 'don't go and X!'. In the same vein, if a speaker wants their addressee to do something at a different location and they explicitly wish to prevent the other option, they might say 'don't X here!'. However, such speakers would be issuing comparatively convoluted directives and would probably be more likely to just say 'come and X!' and 'X there!'.

3.4. Intersubjectivity

As observed in Section 2.4, languages may make more intersubjective distinctions in the negative imperative than in the imperative but, cross-linguistically, neutralization of such marking is clearly more typical from positive to negative than vice versa. The follow-up question in this section, like in Sections 3.2 and 3.3, is whether or not this asymmetry tendency reflects usage at all, in English and Dutch.

The (negative) imperative in these two languages can be modified in a variety of ways to alter its illocutionary strength and/or manage interpersonal relations. Unfortunately, the present article does not have the space to discuss them in any detail.

Some examples in (55) and (56) and some references will therefore have to suffice here (but see Van Olmen 2011: 84–107, 120–127, 135–181). The strategies in English include – inter alia – *do*-support (see Section 3.1), *just* (e.g. Aijmer 2002: 153–174), *please* (e.g. Wichmann 2004), explicating *you* (e.g. De Clerck 2006: 356–397) and tag questions (e.g. Kimps & Davidse 2008), as illustrated in (55a) to (55e) respectively.

(55) a. **Do** hang your coat up if you'd like.
(ICE-GB: S1A.066.7)
b. Now **just** shut up and listen to me.
(ICE-GB: S1A.086.209)
c. Yes **please** don't bother for a moment.
(ICE-GB: S1B.070.138)
d. **You** be careful going back.
(ICE-GB: S1A.019.153)
e. Don't tell **will you**.
(ICE-GB: S1A.032.182)

(56) a. *Maar doe **alsjeblieft** niet meer dan tien.*
'But please don't do more than ten.'
(CGN: fn009146.15)
b. *Let op **hè**.*
'Be careful, won't you.'
(CGN: fn000320.138)
c. *Laat **u** mij nou even uitpraten.*
'You just let me finish talking now.'
(CGN: fn007126.154)
d. *Wees **nou maar** niet zo bang.*
'Just don't be so afraid now.'
(CGN: fn007228.188)
e. *Denk d'r **'ns** over na.*
'Just think about it.'
(CGN: fn007265.138)

The Dutch strategies comprise – among other things – *alsjeblieft* 'please', clause-final particles (e.g. Kirsner 2003) and the formal second person pronoun *u* (e.g. Fortuin

2004), as exemplified in (56a) to (56c) respectively, alongside an array of modal particles (e.g. Vismans 1994), like those in (56d) and (56e).

Usage in the two languages could be said to mirror the cross-linguistic tendency at issue if negative imperatives occurred less often with such intersubjective modification than imperatives. Accordingly, we counted how many (negative) imperatives in Van Olmen's (2011) data are modified. The absolute and relative figures are given in Table 9.

	Speech		**Plays**		**Total**	
	English	**Dutch**	**English**	**Dutch**	**English**	**Dutch**
Imperatives	118 / 738	124 / 250	72 / 596	169 / 288	190 / 1,334	293 / 538
	15.99%	49.60%	12.08%	58.68%	12.24%	54.46%
Negative imperatives	12 / 119	5 / 15	17 / 131	23 / 74	29 / 250	28 / 89
	10.08%	33.33%	12.98%	31.08%	11.60%	31.46%

Table 9: Intersubjective modification in the (negative) imperative in Van Olmen's (2011) corpus data

There is substantially more modification in the imperative than in its negative equivalent in the Dutch plays (χ^2 (df 1) = 18.00 with $p < 0.0001$). In Dutch speech too, we find a higher proportion of modified imperatives but the very low number of negative imperatives makes it impossible to establish a statistically significant difference. We are in a similar position for English, because of its comparatively low rate of modification of (negative) imperatives (ranging from 10.08% to 15.99%, as opposed to 31.08% to 58.68% for Dutch).

For more data, we looked at the TenTen corpora. Our English searches focused on (negative) imperatives that consist of just a verb (e.g. *go!*; *don't go!*; *don't!*) and on verb-only cases that contain *please*, *just* or a tag question (e.g. *go, please!*; *just don't!*; *don't go, will you?*) (see Appendix 2 for the queries).[11] Comparing their frequencies can give us some idea of the degree to which (negative) imperatives have intersubjective modification in the language. For Dutch, we filtered Table 4's dataset of (negative) imperatives for cases that feature one or more of the following items: *alsjeblieft*

[11] Emphatic *do* is not included here because it is not an option in the negative imperative. Explicit *you* is excluded because the 440 hits for our imperative query were rife with false positives (e.g. *you bet!*) and superficially ambiguous hits (e.g. *you decide!*) (note, though, that we only found five cases of *don't you …!*).

and its variants, the clause-final particle *hè* 'will/won't you?' and the (fairly untranslatable) modal particles *dan*, *toch*, *maar*, *eens/'ns*, *even/effe/eventjes, gerust* and *gewoon*.[12] Especially these last words are highly multifunctional and may thus well have a function other than mitigating or reinforcing the (negative) imperative in particular cases (e.g. *even* could still express its original meaning of 'for a short time'). In our view, however, such instances should largely cancel one another out when contrasting the imperative and its negative counterpart. The modal particles also constitute quite a productive category in Dutch and, hence, the present list may not be complete. We believe that it contains the most common ones, though (see Van Olmen 2011: 86–87, 121–122). The corpus findings for both English (.uk) and Dutch (.nl) are presented in Table 10 as the proportions of (negative) imperatives that are modified in absolute and relative terms. Some examples are given in (57).

	English	**Dutch**
Imperatives	634 / 6,994	43,568 / 195,567
	9.06%	22.28%
Negative imperatives	31/ 864	669 / 7,066
	3.59%	9.47%

Table 10: Intersubjective modification in the (negative) imperative in enTenTen20 and nlTenTen20

(57) a. **Please** go!
(enTenTen20: 83105880)

b. *Donder **toch** op met je vliegtuigen!*
'Just fuck off with your airplanes!'
(nlTenTen20: 52544)

c. **Just** don't ASK!
(enTenTen20: 56557670)

d. *Maak mij voor de mensen **toch** niet te schande!*
'Just don't disgrace me before the people!'
(nlTenTen20: 255769)

[12] Table 4's dataset does not cover (negative) imperatives with overt subjects, like (56c), and they are therefore not taken into account here. An additional reason for their exclusion is that they are hard to separate from interrogative clauses (e.g. *ga jij toch weg!* 'you just go away!' versus *ga jij toch weg?* 'are you nevertheless going away?').

The enTenTen20 results suggest that there is a difference in English after all: modified versus bare imperatives occur at a ratio of one to ten while modified versus bare negative imperatives only occur at a ratio of one to 27 (χ^2 (df 1) = 29.78 with $p < 0.00001$). Such a contrast is observed for both *please* (with respective ratios of one to thirteen and one to 35) and *just* (whose respective ratios are one to 50 and one to 119). Our single hit for tag questions occurs after an imperative. The findings from nlTenTen20 confirm those from the Dutch plays in Table 5: the imperative contains intersubjective modification significantly more frequently than its negative equivalent (22.28% versus 9.47%; χ^2 (df 1) = 655.76 with $p < 0.00001$).

In short, there appears to be usage data in English and Dutch supporting the typological tendency that intersubjective distinctions in the imperative disappear in the negative imperative. An obvious question that remains to be answered is why imperative negation exhibits this asymmetry. It might be tempting to invoke the discourse presuppositionality of negation again (see Section 1), as it can explain neutralization in other domains. We are not convinced, however, that it really applies to intersubjectivity in imperative negation or, in other words, that the contextual presence of the positive state of affairs would somehow weaken the wish or requirement to alter illocutionary strength and/or manage rapport in a negative imperative. It is unclear to us, for instance, why an interpersonal relationship that calls for the use of a polite imperative construction in some language/culture does not always create a corresponding "demand" for a polite negative imperative construction. One could possibly counter that the desire or need to get someone to quit doing something or to abstain from an expected course of action supersedes any intersubjective considerations of politeness and mitigation. Then again, this desire or need may equally well be said to motivate the (hypothetical) existence of more peremptory negative imperative than imperative constructions. Furthermore, certain scholars (e.g. De Clerck 2006: 279–282) have in fact argued that negative imperatives are, in general, more face-threatening than their positive equivalents. They risk damaging not only the addressee's "negative face" or "desire to be unimpeded in [their] actions" – like imperatives – but also their "positive face" or desire "to be approved of" (Brown & Levinson 1987: 13), since saying 'don't!' to someone implies a rejection of their current or anticipated conduct. If this argument is correct, it is actually somewhat strange that negative imperatives tend to exhibit fewer means for changing illocutionary strength and/or interpersonal management than imperatives.

It is probably clear from the discussion in the previous paragraph that, at present, we have no real explanation for the facts about intersubjectivity in imperative negation. A very tentative final hypothesis transcends the (negative) imperative and involves the wider range of (negative) directive strategies in a language. The above-mentioned possible difference in face-threatening potential may simply make speakers opt for less established, more novel strategies more often when performing a negative directive speech act than when performing a positive one. These strategies would then serve particular intersubjective purposes but, importantly, they would not necessarily grammaticalize into specialized negative imperative constructions. If they became frequent enough for such a development to take place, they would no longer be "useful": their value as a means to counteract the more serious face threat of a negative directive lies precisely in their lack of conventionality. Such strategies would – and should – not be part of any study of imperative negation proper (see Section 2.1) but their absence might account for its typical asymmetry in intersubjectivity. This suggestion is, of course, highly speculative and will have to remain so here. Support for it could come from research examining and comparing the whole range of positive and negative directive strategies in a variety of languages. This line of investigation is clearly beyond the present article's scope, however.[13]

3.5. Interim summary

The results of this section's corpus studies suggest that the usage of English and Dutch (negative) imperatives indeed reflects the cross-linguistic asymmetries of neutralization from positive to negative in the imperative domain. First, Section 2.2 concludes that, in the languages of the world, tense distinctions between immediate and delayed compliance often disappear from positive to negative but never the other way around. Correspondingly, Section 3.2 shows that the English and Dutch negative imperative tends to feature fewer expressions to do with the time of compliance than its positive equivalent. Second, the evidence in Section 2.3 suggests that, in the world's languages, distinctions of a directional and/or locational nature can be made in both positive and negative imperatives or just in positive ones but never only in negative

[13] Dutch might prove telling, though (see Van Olmen 2010: 478, Devos & Van Olmen 2013: 3–4). It has a history of directive strategies that compete with the negative imperative in particular but disappear fairly quickly. They include *wil niet treuren!* (lit. 'don't want to mourn!'), *niet te treuren!* (lit. 'not to mourn!) and *niet treuren!* ('not mourn!') 'don't mourn!'.

ones. The usage data in Section 3.3 is in line with this cross-linguistic trend, in that the negative imperative in English and Dutch is found to contain fewer directional and/or locational expressions than its positive counterpart. Third, and finally, we argue in Section 2.4 that, while it is possible in language for intersubjective distinctions in the imperative domain to disappear from negative to positive, neutralization in the opposite direction appears to be more common cross-linguistically. Section 3.4 confirms that, in English and Dutch too, the negative imperative features fewer intersubjective expressions than its positive equivalent.

The discourse presuppositionality of negation, invoked before for similar results in standard negation, can be taken as a possible explanation for these usage and typological facts about tense as well as direction and/or location. Yet, we hypothesize that they may also be motivated by more particular factors, such as the less "time-specific" nature of negative imperatives and the comparative inconsequentiality of direction and location in negative directives. Moreover, for the corpus and cross-linguistic results about intersubjectivity, the discourse presuppositionality of negation does not actually appear to be an especially satisfactory explanation. We do not at present have an alternative but believe that it could be fruitful to consider negative directive strategies more generally for an answer.

Of final note is a remarkable parallel between the frequencies with which languages across the world make temporal, directional/locational and intersubjective distinctions in the (negative) imperative and those with which (negative) imperatives in the two languages focused on contain such expressions: intersubjectivity is expressed much more often than tense and direction/location both cross-linguistically and in English and Dutch usage. This similarity is probably not a coincidence, pointing to the relative (un)importance of these distinctions for the imperative domain.

4. Conclusions

In this article, we have tried to respond to Miestamo's (2005) largely unanswered call to extend the study of (a)symmetry to non-declarative negation, by examining a balanced sample of the world's languages for asymmetries in imperative negation concerning three different types of distinctions. We have also attempted to address Miestamo et al.'s (2022) recent programmatic appeal to compare typological findings

with and interpret them in light of usage, by investigating how said distinctions manifest themselves in corpus data on English and Dutch (negative) imperatives. The results of our endeavors have been summarized in detail in Sections 2.5 and 3.5 and, for the sake of conciseness, they will not be repeated here. Instead, we wish to conclude our article with some considerations of a broader nature.

First, widening the study of (a)symmetry's scope to other domains of negation is invaluable, as it deepens our understanding of negation in general, but much work remains to be done in this area (e.g. interrogative negation). For this research, it is important to bear in mind any peculiarities of the domain in question (e.g. intersubjectivity as a dimension relevant to imperative negation; cf. Miestamo & van der Auwera 2007) and that the types of asymmetry known from standard negation may but need not occur in other domains (e.g. the possibility of neutralization from negative to positive here; see also Van Olmen 2024 on finiteness asymmetry in imperative negation).

Second, the relationship between typology and usage deserves to be explored further, for negation as well as for other domains. We are aware that such work should preferably involve more than the two very closely related languages focused on in the present article. This ideal can only become a reality, however, through a concentrated joint effort by numerous linguists. This type of collaboration between people highly familiar with a range of different languages is needed especially because comparing certain expressions' frequencies of occurrence in positive versus negative clauses is just a first step in the study of usage. Our more in-depth analysis of time of compliance, for instance, has revealed an apparent property of negative imperatives that may account for the cross-linguistic tendency to neutralize tense distinctions.

Third, and lastly, general functional explanations for typological tendencies, such as negation's discourse presuppositionality for neutralization from positive to negative, should be attempted and merit (more) serious consideration (than they are occasionally given; e.g. Cristofaro 2021). At the same time, caution is always warranted. They may not stand up to closer scrutiny (e.g. negation's discourse presuppositionality for the neutralization of intersubjective distinctions) and more specific motivations may be available (e.g. the less "time-specific" character of negative imperatives; see also van der Auwera & Devos 2012 on the role of diachrony in (ir)realis marking in imperative negation).

Acknowledgments

A first version of this paper was presented in Helsinki on 20 May 2022 at the *Negation in Clause Combining: Typological and Usage-based Perspectives* workshop organized by Matti Miestamo, Ksenia Shagal, Olli Silvennoinen and Héloïse Calame. Thanks are due to the audience and to two anonymous reviewers for their useful comments.

Abbreviations

1 = 1st person
2 = 2nd person
3 = 3rd person
ACC = accusative
AND = andative
AOR = aorist
APPR = apprehensive
AUX = auxiliary
COMPL = completive
CON = conative
CONNEG = connegative
CONTEMP = contemporative
DAT = dative
DEF = definite
DEL = delayed
DEM = demonstrative
DISLOC = dislocative
DIST = distal
DU = dual
EMP = emphatic
EXCL = exclusive
F = feminine
FAM = familiar
FIN = finite
FUR = further
FUT = future
FV = final vowel
GEN = genitive
IMM = immediate
IMP = imperative
INDF = indefinite
INFML = informal
IPFV = imperfective
IRR = irrealis
LOC = locative
N1 = non-1st person
NEG = negation
NFIN = non-finite
NMLZ = nominalization
NOM = nominative
OBJ = object
OBL = oblique
PFV = perfective
PL = plural
POL = polite
POSS = possessive
PRIV = privative
PROG = progressive
PROH = prohibitive
Q = interrogative
REAL = realis
REDUP = reduplication
REFL = reflexive
REL = relative
REM = remote
SBJV = subjunctive
SEMB = semblative
SG = singular
SS = same subject
VEN = venitive

References

Abdel-Hafiz, Ahmed S. 1988. *A reference grammar of Kunuz Nubian*. Buffalo: State University of New York. (Doctoral Dissertation.)

Aijmer, Karin. 2002. *English discourse particles: Evidence from a corpus*. Amsterdam: John Benjamins.

Aikhenvald, Alexandra Y. 2010. *Imperatives and commands*. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. 2014. On future in commands. In Philippe De Brabanter, Mikhail Kissine & Saghie Sharifzadeh (eds.), *Future times, future tenses*, 205–218. Oxford: Oxford University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Brown, Penelope & Stephen Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Carlin, Eithne B. 2004. *A grammar of Trio: A Cariban language of Suriname*. Frankfurt: Lang.

Cristofaro, Sonia. 2012. Descriptive notions vs. grammatical categories: Unrealized states of affairs and 'irrealis'. *Language Sciences* 34(2). 131–146. https://doi.org/10.1016/j.langsci.2011.08.001

Cristofaro, Sonia. 2021. Towards a source-oriented approach to typological universals. In Peter Arkadiev, Jurgis Pakerys, Inesa Šeškauskienė & Vaiva Žeimantienė (eds.), *Studies in Baltic and other languages: A festschrift for Axel Holvoet on the occasion of his 65th birthday*, 97–117. Vilnius: Vilnius University Press.

Davidson, Matthew. 2002. *Studies in Southern Wakashan (Nootkan) grammar*. Buffalo: University of New York. (Doctoral Dissertation.)

De Clerck, Bernard. 2006. *The imperative in English: A corpus-based pragmatic analysis.* Ghent: Ghent University. (Doctoral Dissertation.)

de Sousa, Hilário. 2006. *The Menggwa Dla language of New Guinea*. Sydney: University of Sydney. (Doctoral Dissertation.)

Devos, Maud & Daniel Van Olmen. 2013. Describing and explaining the variation of Bantu imperatives and prohibitives. *Studies in Language* 37(1). 1–57. https://doi.org/10.1075/sl.37.1.01dev

Donohue, Mark. 1999. *A grammar of Tukang Besi*. Berlin: De Gruyter.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292. https://doi.org/10.1075/sl.13.2.03dry

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68(1). 81–138. https://doi.org/10.2307/416370

Dryer, Matthew S. 2013. Genealogical language list. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://wals.info/languoid/genealogy (Accessed 2023.03.29)

Enrico, John. 2003. *Haida syntax: Volume 1*. Lincoln: University of Nebraska Press.

Evans, Nicholas D. 1995. *A grammar of Kayardild: With historical-comparative notes on Tangkic*. Berlin: De Gruyter.

Fleck, David W. 2003. *A grammar of Matses*. Houston: Rice University. (Doctoral Dissertation.)

Fortescue, Michael. 1984. *West Greenlandic*. London: Croom Helm.

Fortuin, Egbert. 2004. De syntaxis van imperatiefsubjecten en modale partikels: Een pragma-semantische benadering. *Nederlandse Taalkunde* 9(4). 355–375.

Gossner, Jan D. 1994. *Aspects of Edolo grammar*. Arlington: University of Texas. (Doctoral Dissertation.)

Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33. https://doi.org/10.1515/COG.2008.001

Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633. https://doi.org/10.1017/S0022226720000535

Hasselgård, Hilde. 2010. *Adjunct adverbials in English*. Cambridge: Cambridge University Press.

Hyslop, Gwendolyn. 2011. *A grammar of Kurtöp*. Eugene: University of Oregon. (Doctoral Dissertation.)

Jakubícek, Miloš, Adam Kilgarriff, Vojtech Kovár, Pavel Rychlý & Vit Suchomel. 2013. The TenTen corpus family. *Proceedings of the International Corpus Linguistics Conference* 7. 125–127.

Jary, Mark & Mikhail Kissine. 2016. When terminology matters: The imperative as a comparative concept. *Linguistics* 54(1). 119–148. https://doi.org/10.1515/ling-2015-0039

Jung, Ingrid. 2008. *Gramática del páez o nasa yuwe: Descripción de una lengua indígena de Colombia*. Munich: Lincom.

Kimball, Geoffrey D. 1991. *Koasati grammar*. Lincoln: University of Nebraska Press.

Kimps, Ditte & Kristin Davidse. 2008. Illocutionary force and conduciveness in imperative constant polarity tag questions: A typology. *Text & Talk* 28(6). 699–722. https://doi.org/10.1515/TEXT.2008.036

Kirsner, Robert S. 2003. On the interaction of the Dutch pragmatic particles *hoor* and *hè* with the imperative and imperativus pro imperative. In Arie Verhagen & Jeroen van de Weijer (eds.), *Usage-based approaches to Dutch*, 59–96. Utrecht: Netherlands Graduate School of Linguistics.

Langacker, Ronald W. 1990. Subjectification. *Cognitive Linguistics* 1(1). 5–38.

Maslova, Elena. 2003. *A grammar of Kolyma Yukaghir*. Berlin: De Gruyter.

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

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. 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 & Johan van der Auwera. 2007. Negative declaratives and negative imperatives: Similarities and differences. In Andreas Ammann (ed.), *Linguistics festival: May 2006, Bremen*, 59–77. Bochum: Brockmeyer.

Miestamo, Matti, Dik Bakker & Antti Arppe. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296. https://doi.org/10.1515/lingty-2016-0006

Miestamo, Matti, Ksenia Shagal & Olli O. Silvennoinen. 2022. Typology and usage: The case of negation. *Linguistic Typology at the Crossroads* 2(1). 121–154. https://doi.org/10.6092/issn.2785-0943/13508

Miestamo, Matti, Olli O. Silvennoinen & Chingduang Yurayong. 2024. Asymmetry in temporal specification between affirmation and negation: Adverbials and tense-aspect neutralization. *Studies in Language*. aop https://doi.org/10.1075/sl.23036.mie

Morris, Henry F. & Brian E.R. Kirwan. 1972. *A Runyankore grammar*. Nairobi: East African Literature Bureau.

Mourigh, Khalid. 2015. *A grammar of Ghomara Berber*. Leiden: Leiden University. (Doctoral Dissertation.)

Nederlandse Taalunie. 2004. *Corpus gesproken Nederlands.* Release 1.0. The Hague.

Nicolle, Steve. 2009. *Go*-and-V, *come*-and-V, *go*-V and *come*-V: A corpus-based account of deictic movement verb constructions. *English Text Construction* 2(2). 185–208. https://doi.org/10.1075/etc.2.2.03nic

Overall, Simon E. 2017. Commands and prohibitions in Aguaruna. In Alexandra Y. Aikhenvald & R.M.W. Dixon (eds.), *Commands: A cross-linguistic typology*, 61–82. Oxford: Oxford University Press.

Patte, Marie-France. 2008. *Parlons Arawak: Une langue amérindienne d'Amazonie*. Paris: L'Harmattan.

Pensalfini, Robert. 2003. *A grammar of Jingulu, an Aboriginal language of the Northern Territory*. Canberra: Pacific Linguistics.

Rice, Keren. 1989. *A grammar of Slave*. Berlin: De Gruyter.

Rupp, James E. 1989. *Lealao Chinantec syntax*. Dallas: Summer Institute of Linguistics.

Sarvasy, Hannah S. 2017. Imperatives and commands in Nungon. In Alexandra Y. Aikhenvald & R.M.W. Dixon (eds.), *Commands: A cross-linguistic typology*, 224–249. Oxford: Oxford University Press.

Steeman, Sander. 2011. *A grammar of Sandawe: A Khoisan language of Tanzania*. Utrecht: Netherlands Graduate School of Linguistics.

Stevenson, Roland, C. 1969. *Bagirmi grammar*. Khartoum: University of Khartoum.

Survey of English Usage. 2006. *International corpus of English: The British component.* Release 2. London.

Telban, Borut. 2017. Commands as a form of intimacy among the Karawari of Papua New Guinea. In Alexandra Y. Aikhenvald & R.M.W. Dixon (eds.), *Commands: A cross-linguistic typology*, 266–282. Oxford: Oxford University Press.

Traugott, Elizabeth C. 2003. From subjectification to intersubjectification. In Raymond Hickey (ed.), *Motives for language change*, 124–139. Cambridge: Cambridge University Press.

van der Auwera, Johan. 2005. Imperatives. In Keith Brown (ed.), *Encyclopedia of language and linguistics*, 565–567. Amsterdam: Elsevier.

van der Auwera, Johan & Maud Devos. 2012. Irrealis in positive imperatives and in prohibitives. *Language Sciences* 34(2). 171–183. https://doi.org/10.1016/j.langsci.2011.08.003

van der Auwera, Johan & Ludo Lejeune. 2013. The prohibitive. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://wals.info/chapter/71. (Accessed 2023.03.27).

Van der Voort, Hein. 2004. *A grammar of Kwaza*. Berlin: De Gruyter.

Van Olmen, Daniel. 2010. Typology meets usage: The case of the prohibitive infinitive in Dutch. *Folia Linguistica* 44(2). 471–508. https://doi.org/10.1515/flin.2010.017

Van Olmen, Daniel. 2011. *The imperative in English and Dutch: A functional analysis in comparable and parallel corpora*. Antwerp: University of Antwerp. (Doctoral Dissertation.)

Van Olmen, Daniel. 2019. A three-fold approach to the imperative's usage in English and Dutch. *Journal of Pragmatics* 139. 146–162.

https://doi.org/10.1016/j.pragma.2018.11.006

Van Olmen, Daniel. 2021. On order and prohibition. *Studies in Language* 45(3). 520–556. https://doi.org/10.1075/sl.19036.van

Van Olmen, Daniel. 2024. Specialization and finiteness (a)symmetry in imperative negation: With a comparison to standard negation. *Linguistic Typology* 28(2). 205-252. https://doi.org/10.1515/lingty-2022-0007

Van Olmen, Daniel, Maud Devos & Valentin Rădulescu. 2023. (A)symmetries in imperative negation in Eastern Bantu. *Africana Linguistica* 29. 179–222.

Vismans, Roel. 1994. *Modal particles in Dutch directives: A study in functional grammar.* Dordrecht: ICG Printing.

Vuillermet, Marine. 2012. *A grammar of Ese Ejja, a Takanan language of the Bolivian Amazon*. Lyon: Lumière University of Lyon 2. (Doctoral Dissertation.)

West, Birdie. 1980. *Gramatica popular del Tucano*. Bogotá: Linguistics Institute of Verano.

Wichmann, Anne. 2004. The intonation of *please*-requests: A corpus-based study. *Journal of Pragmatics* 36(9). 1521–1549. https://doi.org/10.1016/j.pragma.2004.03.003

Wilbur, Joshua. 2014. *A grammar of Pite Saami*. Berlin: Language Science Press.

Xrakovskij, Viktor S. (ed.). 2001. *Typology of imperative constructions*. Munich: Lincom.

CONTACT

d.vanolmen@lancaster.ac.uk

Appendix 1

Find below the following information on each of the 160 languages in the sample used for Section 2: its macroarea, language, genus, Glottolog code and ISO 639-3 code.

Macroarea	Language Family	Genus	Language	Glottolog	ISO 639-3
Africa	Afro-Asiatic	Lowland East Cushitic	Somali	soma1255	som
		North Omotic	Wolaitta	wola1242	wal
		Semitic	Egyptian Arabic	egyp1253	arz
		West Chadic	Hausa	haus1257	hau
	Central Sudanic	Kresh	Kresh	gbay1288	krs
	Dogon	Dogon	Penange	pena1270	n/a
	Eastern Sudanic	Kuliak	So	sooo1256	teu
		Nilotic	Lango	lang1324	laj
		Nubian	Kunuz Nubian	kenu1236	kzh
	Gumuz	Gumuz	Northern Gumuz	gumu1244	guk
	Kadu	Kadulgi	Krongo	kron1241	kgo
	Khoe-Kwadi	Khoe-Kwadi	Nama	nama1264	naq
	Koman	Koman	Komo	komo1258	xom
	Kxa	Ju-Kung	Ju|'hoan	juho1239	ktz
	Maban	Maban	Maba	maba1277	mde
	Mande	Eastern Mande	Busa	busa1253	bqp
		Western Mande	Jalkunan	jalk1242	bxl
	Niger-Congo	Bantoid	Shangaci	nath1238	nte
		Defoid	Yoruba	yoru1245	yor

		Edoid	Degema	dege1246	deg
	Saharan	Western Saharan	Kanuri	cent2050	knc
	Sandawe	Sandawe	Sandawe	sand1273	sad
	Songhay	Songhay	Koyraboro Senni	koyr1242	ses
Australia & New Guinea	Anim	Marind	Marind	hali1245	mrz
	Border	Border	Imonda	imon1245	imn
	Bosavi	Bosavi	Edolo	edol1239	etr
	Dagan	Dagan	Daga	daga1275	dgz
	Darwin Region	Laragia	Laragia	lara1258	lrg
	Eleman	Tate	Kaki Ae	kaki1249	tbd
	Gaagudju	Gaagudju	Gaagudju	gaga1251	gbu
	Garrwan	Garrwan	Garrwa	gara1269	wrk
	Iwaidjan	Iwaidjan	Maung	maun1240	mph
	Kolopon	Kolopon	Kimaghama	kima1246	kig
	Lower Sepik-Ramu	Lower Sepik	Karawari	tabr1243	tzx
	Mangarrayi-Maran	Mangarrayi	Mangarrayi	mang1381	mpc
	Mangrida	Burarran	Gurr-goni	gura1251	gge
	Mirndi	Djingili	Djingili	djin1251	jig
		Wambayan	Wambaya	wamb1258	wmb
	Morehead and Upper Maro Rivers	Morehead and Upper Maro Rivers	Komnzo	wara1294	tci
	Pama-Nyungan	Northern Pama-Nyungan	Yidiny	yidi1250	yii
		Southeastern Pama-Nyungan	Ngiyambaa	wang1291	wyb
		Western Pama-Nyungan	Ritharngu	rita1239	rit
	Senagi	Senagi	Menggwa	dera1245	kbv
	Sentani	Sentani	Sentani	nucl1632	set
	Sepik	Middle Sepik	Manambu	mana1298	mle

		Sepik Hill	Alamblak	alam1246	amp
	Solomons East Papuan	Lavukaleve	Lavukaleve	lavu1241	lvk
		Savosavo	Savosavo	savo1255	svs
	South Bird's Head	Inanwatan	Inanwatan	suab1238	szp
	Sulka	Sulka	Sulka	sulk1246	sua
	Tangkic	Tangkic	Kayardild	kaya1319	gyd
	Timor-Alar-Pantar	Greater Alor	Adang	adan1251	and
		Makasae-Fataluku-Oirata	Makalero	maka1316	mkz
	Tiwian	Tiwian	Tiwi	tiwi1244	tiw
	Torricelli	Marienberg	Kamasau	kama1367	kms
	Trans-New Guinea	Madang	Kobon	kobo1249	kpw
		Asmat-Kamoro	Asmat	cent2117	cns
		Awju-Dumut	Korowai	koro1312	khe
		Binanderean	Suena	suen1241	sue
		Finisterre-Huon	Nungon	yaum1237	yuw
	Wagiman	Wagiman	Wagiman	wage1238	waq
	West Bomberai	West Bomberai	Kalamang	kara1499	kgv
	West Papuan	North-Central Bird's Head	Abun	abun1252	kgr
	Western Daly	Wagaydy	Emmi	amii1238	amy
	Worrorran	Worrorran	Gunin	kwin1241	gww
	Yareban	Yareban	Yareba	yare1248	yrb
Eurasia	Altaic	Tungusic	Evenki	even1259	evn
	Basque	Basque	Basque (Western)	basq1248	eus
	Burushaski	Burushaski	Burushaski	buru1296	bsk
	Chukotko-Kamchatkan	Southern Chukotko-Kamchatkan	Itelmen	itel1242	itl
	Dravidian	Northern Dravidian	Brahui	brah1256	brh

	Indo-European	Germanic	Icelandic	icel1247	isl
	Japonic	Japonic	Japanese	nucl1643	jpn
	Nahali	Nahali	Nahali	niha1238	nll
	Nakh-Daghestanian	Lezgic	Lezgian	lezg1247	lez
	Nivkh	Nivkh	Nivkh	nivk1234	niv
	Uralic	Saami	Pite Saami	pite1240	sje
	Yenesian	Yenesian	Ket	kett1243	ket
	Yukaghir	Yukaghir	Kolyma Yukaghir	sout2750	yux
North America	Algic	Algonquian	Plains Cree	plai1258	crk
	Caddoan	Caddoan	Wichita	wich1260	wic
	Eskimo-Aleut	Eskimo	West Greenlandic	kala1399	kal
	Haida	Haida	Haida	haid1248	hai
	Hokan	Pomoan	Southern Pomo	sout2984	peq
		Yuman	Maricopa	mari1440	mrc
	Iroquoian	Northern Iroquoian	Oneida	onei1249	one
	Keresan	Keresan	Acoma	west2632	kjq
	Kiowa-Tanoan	Kiowa-Tanoan	Kiowa	kiow1266	kio
	Kutenai	Kutenai	Kutenai	kute1249	kut
	Mayan	Mayan	Mam	mamm1241	mam
	Mixe-Zoque	Mixe-Zoque	Zoque (Copainalá)	copa1236	zoc
	Muskogean	Muskogean	Koasati	koas1236	cku
	Oto-Manguean	Chinantecan	Chinantec Lealao	leal1235	cle
		Popolocan	Mixtec Chalca-tongo	sanm1295	mig
	Penutian	Sahaptian	Nez Perce	nezp1238	nez
		Wintuan	Wintu	wint1259	wit

	Salishan	Interior Salish	Shuswap	shus1248	shs
	Siouan	Core Siouan	Lakota	lako1247	lkt
	Tarascan	Tarascan	Purépecha	pure1242	tsz
	Tonkawa	Tonkawa	Tonkawa	tonk1249	tqw
	Totonacan	Totonacan	Huehuetla Tepehua	hueh1236	tee
	Uto-Aztecan	Aztecan	Nahuatl Tetelcingo	tete1251	nhg
		Numic	Northern Paiute	nort2954	pao
	Wakashan	Southern Wakashan	Nuuchahnulth	nuuc1236	nuk
	Wappo-Yukian	Wappo	Wappo	wapp1239	wao
	Yuchi	Yuchi	Yuchi	yuch1247	yuc
	Zuni	Zuni	Zuni	zuni1245	zun
South America	Andoke	Andoke	Andoke	ando1256	ano
	Arauan	Arauan	Paumarí	paum1247	pad
	Araucanian	Araucanian	Mapudungun	mapu1245	arn
	Aymaran	Aymaran	Jaqaru	jaqa1244	jqr
	Barbacoan	Barbacoan	Awa Pit	awac1239	kwi
	Bororoan	Bororoan	Bororo	boro1282	bor
	Cahuapanan	Cahuapanan	Jebero	jebe1250	jeb
	Cariban	Cariban	Trio	trio1238	tri
	Chapacura-Wanham	Chapacura-Wanham	Wari'	wari1268	pav
	Chibchan	Rama	Rama	rama1270	rma
	Choco	Choco	Epena Pedee	epen1239	sja
	Guahiban	Guahiban	Cuiba	cuib1242	cui
	Huitotoan	Huitoto	Murui	muru1274	huu
	Jivaroan	Jivaroan	Aguaruna	agua1253	agr
	Kwaza	Kwaza	Kwazá	kwaz1243	xwa

	Matacoan	Matacoan	Chorote	iyow1239	crq
	Mosetenan	Mosetenan	Mosetén	mose1249	cas
	Mura	Mura	Pirahã	pira1253	myp
	Nadahup	Nadahup	Yuhup	yuhu1238	yab
	Páezan	Páezan	Páez	paez1247	pbb
	Panoan	Panoan	Matsés	mats1244	mcf
	Quechuan	Quechuan	Quecha Imbabura	imba1240	qvi
	Sáliban	Sáliban	Mako	maco1239	wpc
	Tacanan	Tacanan	Ese Ejja	esee1248	ese
	Trumai	Trumai	Trumai	trum1247	tpy
	Tucanoan	Tucanoan	Tuyuca	tuyu1244	tue
	Tupian	Tupi-Guaraní	Emerillon	emer1243	eme
	Uru-Chipaya	Uru-Chipaya	Chipaya	chip1262	cap
	Waorani	Waorani	Waorani	waor1240	auc
	Warao	Warao	Warao	wara1303	wba
	Yanomam	Yanomam	Sanuma	sanu1240	xsi
	Yaruro	Yaruro	Yaruro	pume1238	yae
	Yuracare	Yuracare	Yuracare	yura1255	yuz
South East Asia & Oceania	Austro-Asiatic	Aslian	Semelai	seme1247	sza
		Katuic	Pacoh	paco1243	pac
		Khasian	Khasi	khas1269	kha
		Khmer	Khmer	cent1989	khm
		Palaungic	Wa	para1301	prk
	Austronesian	Celebic	Tukang Besi	tuka1248	khc
		Central-Malayo-Polynesian	Kambera	kamb1299	xbr
		North Borneo	Begak	idaa1241	dbj

		North-West Sumatra Barrier Island	Batak Karo	bata1293	btx
		Oceanic	Vitu	mudu1242	wiv
		Paiwan	Paiwan	paiw1248	pwn
	Great Adamanese	Great Adamanese	Great Andamanese	akaj1239	akj
	Hmong-Mien	Hmong-Mien	White Hmong	hmon1333	mww
	Sino-Tibetan	Bodic	Kurtöp	kurt1248	xkz
		Burmese-Lolo	Burmese	nucl1310	mya
		Kuki-Chin	Daai Chin	daai1236	dao
		Qiangic	Qiang	sout2728	qxs
		Sinitic	Cantonese	yuec1235	yue
	Tai-Kadai	Kadai	Zoulei	aoua1234	aou
		Kam-Tai	Lao	laoo1244	lao

Appendix 2

Find below the enTenTen20 queries that we conducted to determine the frequency of intersubjective marking in English (negative) imperatives for Section 3.4.

(i) English bare imperatives

<s> [tag="VV.*|VB.*|VH.*" & word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.*iNG"] [word="\!"] </s>

(ii) English bare negative imperatives

<s> [word="do|DO|Do|dO"] [lemma="not"] [tag="VV.*|VB.*|VH.*" & word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.*iNG"]? [word="\!"] </s>

(iii) English imperatives with *please* or *just*

<s> [lemma="please|just"] [word="\,"]? [tag="VV.*|VH.*|VB.*" & word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.*iNG"] [word="\!"] </s>

<s> [tag="VV.*|VH.*|VB.*" & word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.*iNG"] [word="\,"]? [lemma="please"] [word="\!"] </s>

(iv) English negative imperatives with *please* or *just*

<s> [lemma="please|just"] [word="\,"]? [word="DO|Do|dO|do"] [lemma="not"] [tag="VV.*|VH.*|VB.*" & word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.*iNG"]? [word="\!"] </s>

<s> [word="DO|Do|dO|do"] [lemma="not"] [tag="VV.*|VH.*|VB.*" &
word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|
DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.
*iNG"]? [word="\,"]? [lemma="please"] [word="\!"] </s>

<s> [word="DO|Do|dO|do"] [lemma="not"] [word="\,"]?
[lemma="please|just"] [word="\,"]? [tag="VV.*|VH.*|VB.*" &
word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|
DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.
*iNG"] [word="\!"] </s>

(v) English imperatives with tag questions

<s> [tag="VV.*|VH.*|VB.*" &
word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|
DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.
*iNG"] [word="\,"]? [lemma="will|would|can|could"] [lemma="not"]?
[lemma="you"] [lemma="not"]? [word="\!|\?"]</s>

(vi) English negative imperatives with tag questions

<s> [word="DO|Do|dO|do"] [lemma="not"] [tag="VV.*|VH.*|VB.*" &
word!="done|Done|dOne|doNe|donE|DOne|DoNe|DonE|dONe|dOnE|doNE|DONe|
DoNE|dONE|DONE|.*ED|.*ed|.*Ed|.*eD|.*ING|.*ing|.*Ing|.*iNg|.*inG|.*INg|.*InG|.
*iNG"]? [word="\,"]? [lemma="will|would|can|could"] [lemma="not"]?
[lemma="you"] [lemma="not"]? [word="\!|\?"] </s>

Linguistic Typology at the Crossroads 4-2 (2024): 220-272 ISSN 2785-0943

Noun juxtaposition for predication, possession, and conjunction: Beyond ambiguity avoidance

SHOGO MIZUNO
LEIPZIG UNIVERSITY & KYOTO UNIVERSITY

Submitted: 15/10/2023 Revised version: 13/12/2024
Accepted: 13/12/2024 Published: 23/01/2025

Abstract

This paper asks whether ambiguity avoidance influences the use of certain linguistic forms, using noun juxtaposition as a case study. Noun juxtaposition is one of the strategies for expressing predication, possession, and conjunction, and is widely used across the world's languages. Despite its extensive use, few studies have investigated noun juxtaposition cross-linguistically. One notable exception is Frajzyngier et al. (2002), who argue that the use of noun juxtaposition is constrained within a single language due to ambiguity avoidance. However, counterexamples to this hypothesis exist. This study points out that their sample is skewed towards African languages, and thus, their findings likely reflect African areal patterns. From this perspective, a comprehensive cross-linguistic examination of noun juxtaposition is still lacking. Therefore, based on a balanced 72-language sample, this paper explores which functions tend to be expressed by noun juxtaposition, whether there are any areal patterns concerning its use, and whether its use is constrained by ambiguity. Since noun juxtaposition is, by definition, the most efficient strategy for these three functions in terms of formal complexity, the research conducted in this study contributes to discussions on whether ambiguity or efficiency is more important for the use of certain linguistic forms. Based on the empirical findings, this study suggests that efficiency plays a more important role than ambiguity.

Keywords: noun juxtaposition; ambiguity; efficiency; predication; possession; conjunction

1. Introduction

Ambiguity and efficiency are important factors in explaining the use of certain linguistic forms. However, they can be opposed to each other. The more efficient a form is, the more ambiguous it tends to be. In this paper, I investigate which is more important for the use of a certain linguistic structure: ambiguity or efficiency, through the examination of noun juxtaposition as a case study. In terms of formal complexity, noun juxtaposition can be considered the simplest (and most efficient) form for expressing meanings conveyed by noun phrases, as it does not use any formal markings to indicate its function. In this sense, the examination of noun juxtaposition is well-suited to the discussion of whether ambiguity or efficiency is more significant.

Noun juxtaposition is one of the strategies (in the sense of Croft 2022 and Haspelmath 2024a) for predication, adnominal possession, and conjunction, and it also serves other functions (see Section 3 for the scope of the survey in the present paper). While it is not found in all the world's languages, this strategy is attested in many languages worldwide. This is illustrated in examples (1)-(3), which are drawn from six macroareas.

(1) Predication[1]

a. Warao (wba; Isolate, South America, Guyana; Romero-Figueroa 1997: 11)
yatu *hotarao*
you non.Warao
'You are non-Warao.'

b. Jaminjung (djd; Mirndi, Yirram; Australia; Schultze-Berndt 2000: 109)
ngayug *gurrany* *gujarding* *ngunggina*
1SG NEG mother 2SG.POSS
'I am not your mother.'

(2) Possession

a. Tommo So (dto; Dogon, Escarpment Dogon; Mali; McPherson 2013: 191)
bé *nínɛ́*
they aunt
'their aunt'

[1] In this paper, the notation and glosses of examples may differ from those in their original sources. All information on the geographical and genealogical distribution of languages is based on Glottolog 5.0

b. Haida (hai; Isolate, North America, Canada; Enrico 2003: 709)
Joe 7isgyaan Bill 7aww
Joe and Bill mother.SG
'Joe's and Bill's mother'

(3) Conjunction
a. Ulwa (yla; Keram, Ulmapo; Papua New Guinea; Barlow 2023: 354)
yeta yena ala
man woman PL.DIST
'the boys and girls'

b. Dolgan (dlg; Turkic, Common Turkic; Russian Federation; Däbritz 2022: 320)
n'el'ma-lar, muksu:t-tar, o:mul-lar
nelma-PL muksun-PL Arctic.cisco-PL
'nelmas, muksuns and Arctic ciscos (fish names)'

Although noun juxtaposition is used worldwide, it has not been extensively investigated cross-linguistically. A notable exception is Frajzyngier et al. (2002), who examine the predicational and possessive functions of noun juxtaposition and argue that its use within a single language is constrained by ambiguity. However, as mentioned in Section 2, since the sample of languages in their study is skewed towards African languages, their investigation is not truly worldwide. Thus, it remains to be explored which functions tend to be expressed by noun juxtaposition, whether there are any areal patterns concerning its use, and whether the claim that the use of noun juxtaposition is constrained by ambiguity is supported. In this paper, I present an examination of noun juxtaposition across the world's languages by investigating a balanced sample of 72 languages. Based on the results, I conclude that the use of noun juxtaposition is not constrained by ambiguity. Instead, these results suggest that human languages tend to prioritize efficiency over avoiding ambiguity. This conclusion offers empirical support for the claims made by Piantadosi et al. (2012) and Wasow (2015). As they claim, ambiguity is not always avoided, and the importance of ambiguity avoidance can sometimes be overrated.

2. Previous studies

As mentioned in Section 1, few studies examine whether there are cross-linguistic

tendencies in the use of noun juxtaposition, even though it is employed for a few functions widely. One notable exception is Frajzyngier et al. (2002).

Frajzyngier et al. (2002: 155) argue that a language does not allow the systematic use of the same formal niche for different functions, that is, a language does not allow systematic ambiguity of grammatical constructions. For the purposes of examining their larger claim, they investigate two functions that can be expressed by noun juxtaposition, namely equational predication and modification of one noun by another.[2] They conclude that (i) if equational predication in the unmarked present tense is coded by noun juxtaposition, modification requires a marker, and (ii) if modification is coded by noun juxtaposition, equational predication requires a marker. The following examples, (4) and (5), from Hdi (xed; Afro-Asiatic, Chadic) and Mupun (sur; Afro-Asiatic, Chadic) instantiate (i) and (ii), respectively, with the markers indicated in bold.

(4) Hdi (Afro-Asiatic, Chadic; Frajzyngier et al. 2002: 165)

a. Equational clause

m̀nd-á *ráyá* *mbítsá*

man-GEN hunt Mbitsa

'Mbitsa is a hunter.'

b. Modification

*hlúwí-**á*** *krì*

meat-GEN dog

'dog meat'

(5) Mupun (Afro-Asiatic, Chadic; Frajzyngier et al. 2002: 162–163)

a. Equational clause

wur ***a*** *wat*

he COP thief

'He is a thief.'

[2] As is evident from (4b) and (5b), what they refer to as modification is, in fact, adnominal possession. I follow their use of this terminology when reviewing their study.

b. Modification
siwol *laa*
money child
'child's money'

However, this observation is problematic. In fact, counterexamples to their claim are found in some languages. For example, Sentani (set; Sentanic, Nuclear Sentanic) and Labwor (lth; Nilotic, Western Nilotic) use noun juxtaposition for both functions, as in (6) and (7).

(6) Sentani predication and modification (Sentanic, Nuclear Sentanic; Mayer 2021: 63)
Awansi *Jacobus* *mænggə* *fa.*
Awansi Jacobus girl young
'Awansi is Jacobus's daughter.'

(7) Labwor (Nilotic, Western Nilotic; Heine & König 2010: 30; 61)
a. Predication
mánón *bɔ̀ɔ́*
that bɔɔ
'It is bɔɔ vegetable.'

b. Modification
ɔ̀t *dhákɔ́*
house woman
'woman's house'

Frajzyngier et al.'s (2002) hypothesis is mainly based on African languages, particularly Chadic languages, as their sample includes 11 African languages out of a total of 33 languages. This is why their claim is biased toward African areal patterns and does not work cross-linguistically (see also Kazama 2011 for a critique of Frajzyngier et al. 2002).

Thus, while Frajzyngier et al. (2002) claim that the use of noun juxtaposition is motivated or constrained by ambiguity, it remains largely unexplored whether this claim holds cross-linguistically. Consequently, questions arise as to whether there are

distributional tendencies in the use of noun juxtaposition across languages and areas, and if such tendencies exist, whether they can be explained in terms of ambiguity or efficiency. This paper aims to answer these questions. The next section is dedicated to the preparation for the survey.

3. Definition of terms and the scope of the study

3.1. Noun juxtaposition

To conduct typological research on noun juxtaposition, we need to define it as a comparative concept (Haspelmath 2007a, 2010). In the present study, *noun juxtaposition* is defined as in (8).

(8) Noun juxtaposition
Noun juxtaposition is a structure in which two (or more) nouns occur adjacent to each other in a single construction, and neither of the nouns is marked by a formal marker that indicates a relationship between them.[3]

This definition requires three comments on *noun*. First, it is generally not straightforward how nouns can be compared cross-linguistically, because different languages have different word classes (Evans 2000). In this paper, *noun* is treated as part of universally available concepts (Haspelmath 2023a: 23), as defined in (9).

(9) Noun (Croft 1991: ch.2, 2000, 2001: ch.2, 2022: 714; Haspelmath 2023a)
A noun is a word that is used as an argument of a verb, that is, the head of a referring phrase, and it denotes an object without any additional markers.

This definition of *noun* singles out only typical nouns. Of course, other semantic classes, such as action and property can form nouns, but they need additional markers in many cases (e.g., *walk-walking; kind-kindness*). However, this paper does not address such nouns that require additional markers.

Second, this paper addresses structures in which at least one of the elements involves noun phrases (hereafter referred to as NP). As I mentioned earlier, this paper

[3] I name such formal markers *function indicators*.

investigates whether ambiguity plays a role in explaining the use of certain linguistic forms through the examination of noun juxtaposition, as argued by Frajzyngier et al. (2002). The structures that they examined involve at least one NP as an element. For example, predication involves two NPs, such as *[My mother]*$_{NP}$ *is [her teacher]*$_{NP}$, and possession involves at least one NP, such as in *[his father]*$_{NP}$*'s house*.

Third, this paper includes pronouns within its scope (e.g. (1), (2a), (5a), and (7a), among others). This is because investigating pronouns is also useful for achieving our aims, such as examining which functions are typically expressed by noun juxtaposition, whether there are any areal patterns regarding its use, and whether its use is constrained by ambiguity.

In (8), noun juxtaposition is defined as one of the strategies used to express certain functions (see Croft 2022 and Haspelmath 2024a for the distinction between strategies and functions). One of the aims of the present paper, as mentioned earlier, is to investigate which functions are typically expressed by noun juxtaposition. Therefore, it is important to determine which functions to focus on in this study. Noun juxtaposition can be used not only for predication and possession but also for coordination and other functions, such as apposition. However, this paper focuses only on predication, possession, and conjunction. This is because these functions are often expressed by noun juxtaposition, as mentioned in the following subsections, and there is also potential ambiguity between them. Similar to predication and possession, conjunction involves two NPs as well, such as *[my sister]*$_{NP}$ *and [her brother]*$_{NP}$.[4] Before looking at these three functions in detail, I make six comments on the scope of the survey and explain why functions other than predication, possession, and conjunction are excluded.

First, as I mentioned earlier, this paper focuses on noun phrases and excludes clauses from consideration. Therefore, juxtaposed clauses, such as complementation in Thai (tha; Thai-Kadai, Kam-Tai; Iwasaki & Ingkaphirom 2005: 253–255) are beyond the scope of this study.

Second, the present study does not deal with noun modifiers. This is because there are languages in which nouns and adjectives are not distinguished by morphosyntactic criteria (Plungian 2011: 75). For instance, Huallaga Quechua (qub; Quechuan, Quechua I) does not differentiate between nouns and adjectives

[4] A reviewer questions whether there is ambiguity between clauses and phrases, but it is indeed reported in several grammars. For example, in Sentani, noun juxtaposition is ambiguous in its interpretation between predication and adnominal possession (Mayer 2021: 64).

morphosyntactically, as illustrated in (10). Therefore, all of its property-modificational constructions could fall under the scope if noun modifiers were taken into account (this is relevant to the definition of noun above).

(10) Huallaga Quechua (Quechuan, Quechua I; Weber 1989: 36)
a. *rumi* *wasi*
stone house
'stone house'

b. *hatun* *wasi*
big house
'big house'

Thus, this paper excludes noun modifiers, such as (11a), an example from Araona (aro; Pano-Tacanan, Tacanan), and so-called generic-specific construction such as (11b), an example from Kayardild (gyd; Tangkic, Southern Tangkic) even if noun juxtaposition is used. In Araona, juxtaposed nouns express several meanings other than possessive, such as modification (see Emkow 2006: ch. 13.7.4), and in Kayardild, a generic noun naming a class or use of entities and a specific noun are juxtaposed (see Evans 1995: ch. 6.3.4).

(11) a. Araona (Pano-Tacanan, Tacanan; Emkow 2006: 381)
nāi *bēne*
rain side
'rain side'

b. Kayardild (Tangkic, Southern Tangkic; Evans 1995: 244)
wanku-ya *kulkiji-y*
elasmobranch-LOC shark-LOC
'a shark'

Third, the present study excludes *apposition* from consideration. This is because almost all languages can use the juxtaposition strategy for apposition to some extent (see Hackstein 2003 for the definition of apposition and Logvinova 2024 for the relationship between apposition and juxtaposition). For example, Russian (rus; Indo-

European, Balto-Slavic) and Japanese (jpn; Japonic, Japanesic) are typically regarded as languages in which noun juxtaposition is rarely used except for predication. However, these languages can also use it for apposition, as in (12). Thus, the use of noun juxtaposition for apposition does not seem to be theoretically motivated or constrained.

(12) a. Japanese (Japonic, Japanesic)
Nihon=*no* *syuto* *Tokyoo*=*ni* *ik-u.*
Japan=GEN capital Tokyo=ALL go-NPST
'I will go to Tokyo, the capital of Japan.'

b. Russian (Indo-European, Balto-Slavic; Timberlake 2004: 152)
Ozero *Bajkal* *gluboko.*
lake.N.SG Baikal.M.SG deep
'Lake Baikal is deep.'

Fourth, the present study excludes *compounding* from the scope of the survey. In some languages, possessive compounds and conjunctive compounds (co-compounds in Wälchli 2005) are formed, as possessive compounds in (13).

(13) Bunaq (bfn; Timor-Alor-Pantar, Bunak; Schapper 2022: 350)
deu *puqup*
house roof
'house roof'

This study excludes compounds from the scope of the investigation because compounding involves only Ns, not NPs according to the definition of *compound* proposed by Haspelmath (2023c).

(14) Compound (Haspelmath 2023c: 288)
A compound is a form (consisting of two adjacent roots) that instantiates, or was created by a compound construction, namely, a construction consisting of two strictly adjacent slots for roots that cannot be expanded by full nominal, adjectival, or degree modifiers).

At this point, there is no potential ambiguity between compounding and the three functions in question. Thus, compounding does not contribute to the discussions about whether ambiguity plays a role in the use of certain forms, which are explored in this study.

Fifth, in many cases, the use of noun juxtaposition is restricted to certain conditions, and thus, strategies other than noun juxtaposition can be used in a similar (or the same) way. For example, Yélî Dnye (yle; Isolate, Papunesia) uses a comitative case for conjunction in addition to the noun juxtaposition strategy, as illustrated in (15).

(15) Yélî Dnye (Isolate, Papunesia; Levinson 2022: 163)

a. *Yidika* *Mwonî*
Yidika Mwonî
'Yidika and Mwonî'

b. *Yidika* *Mwonî* *k:ii*
Yidika Mwonî COM
'Yidika and Mwonî'

In this paper, noun juxtaposition is considered to be used in a language if it is employed under certain conditions. I do not investigate the specific conditions under which noun juxtaposition can be used or the difference in semantics and/or information structure between noun juxtaposition and non-juxtaposition strategy. This is because the primary purpose of this study is to examine the relationships between functions, rather than within a single function.

Sixth, in this paper, I do not consider intonation and/or other phonological means. Indeed, such phonological means might be a function indicator in noun juxtaposition. However, as Mithun (1988: 357) notes regarding coordination, the intonational linking of concepts can be universal in spoken language. In addition, phonological effects are quite diverse and cannot be easily generalized across languages (Haspelmath 2023b). Therefore, they are excluded from the scope of this study.

In the following subsections, I examine three functions, namely, predication, adnominal possession, and conjunction which are the focus of this study in detail.

3.2. Predication

In many languages, nouns in juxtaposition can express a predicational relationship.

For example, Kalamang (kgv; West Bomberai, Kalamang), Kugu Nganhcara (uwa; Pama-Nyungan, Paman), and Duhumbi (cvg; Sino-Tibetan, Kho-Bwa) are among such languages, as in (16).

(16) a. Kalamang (West Bomberai, Kalamang; Visser 2022: 293)
kon se guru, tumtum kon guru
one already teacher children one teacher
'One is already teacher, one child is teacher.'

b. Kugu Nganhcara (Pama-Nyungan, Paman; Smith & Johnson 2000: 389)
iiru thata
this.ABS frog
'This is a frog.'

c. Duhumbi (Sino-Tibetan, Kho-Bwa; Bodt 2020: 396)
otɕʰi ɕoj Pema-a?
this bull pema-GEN
'This bull is Pema's.'

In the literature, various subfunctions of predication are distinguished. For example, Haspelmath (2024b) introduces the neologism *duonominal construction* and subdivides it into two types, namely equational clauses and classificational clauses. On the other hand, Croft (2022: ch. 10.1) distinguishes predicational and identificational clauses, based on Stassen (1997: ch. 3.6).[5] Actually, concerning the terms *predication* and *predicational nominal* that have been used in this paper so far, there are cases where they should be regarded as *identification* rather than predication. In many cases, the juxtaposition strategy is used for all subfunctions of predication. However, there are a few languages that use the juxtaposition strategy for only one of these subfunctions. This is the case with Yuchi (yuc; Isolate, North America), where only equational clauses use juxtaposition, as illustrated in (17).

[5] He further subdivides the identificational clause into presentational and equational clauses.

(17) Yuchi (Isolate, North America)

a. Equational clause (Linn 2001: 416)
Josephine senõ se-laga.
Josephine NC:F 3F.POSS-grandmother
'Josephine is her grandmother.'

b. Classificational clause (Linn 2001: 415)
Simon 'wa p'athl'ẽ.
Simon COP chief
'Simon is chief.'

In this paper, I do not distinguish types of predication, such as equational and classificational, and instead use the cover term *predication* for them. This is because the distinctions among these subfunctions vary from one linguist to another, and there is no consensus on the matter. For example, Haspelmath (2024b) makes a distinction between types of predication based on form, while Croft (2022) and Stassen (1997) base their distinctions on cognitive differences (mental-files). Many other proposals (e.g., Payne 1997: ch. 6) have also been made (see Haspelmath 2024b for a summary of the literature). Since the present study does not pursue an appropriate distinction between subfunctions within a single function, such as predication, and instead examines the relationship between use of noun juxtaposition for several functions, a strict distinction between subfunctions within a function is not required. Thus, if a language uses noun juxtaposition for any subfunction of predication, regardless of the type, I consider this language as one that can use the juxtaposition strategy for predication.

The definition of *noun juxtaposition* employed in this paper excludes nouns in the so-called predicative form from noun juxtaposition because they indicate a predicational relationship. Thus, the predicative noun in (18) from Kolyma Yukaghir (yux; Yukaghir, Kolymic) is not counted as an element consisting of noun juxtaposition.

(18) Kolyma Yukaghir (Yukaghir, Kolymic; Maslova 2003: 437)
Momušā laqidī'e čistē čumu amun-ek.
Momusha tail entirely all bone-PRED
'The whole tail of Momusha is only bones.'

In some languages, predicative nouns are regarded as stative verbs because of their predicational function. This is the case with the predicative noun ∅-*kʷɜbʒɜ́* in (19) from Ubykh (uby; Abkhaz-Adyge, Ubykh).

(19) Ubykh (Abkhaz-Adyge, Ubykh; Fenwick 2011: 155)
ɒ-ʒʷɜnkʲɬ́ ∅-*kʷɜbʒɜ́*
the-flea 3SG.ABS-man
'The flea is a male.'

However, it falls under the definition of a noun provided in (9). Thus, (19) can be considered an example of noun juxtaposition in a cross-linguistic context.

3.3. Adnominal possession

Noun juxtaposition can express an adnominal possessive relationship. For example, Ju|'hoan (kyz; Kxa, Ju-Kung), Amur Nivkh (niv; Nivkh, Amur Nivkh), and Rama (rma; Chibchan, Core Chibchan) can use it to express adnominal possession, as illustrated in (20).

(20) a. Ju|'hoan (Kxa, Ju-Kung; Dickens 1992: 17)
n!hai *!xúí*
lion tail
'the lion's tail'

b. Amur Nivkh (Nivkh, Amur Nivkh; Nedjalkov & Otaina 2013: 9)
ətək *χaj*
father pigeon
'father's pigeon'

c. Rama (Chibchan, Core Chibchan; Grinevald 1990: 94)
tangkit *(aing)* *ariira*
bow (POSS) string
'the string of the bow'

In the present study, a *possessive construction* is defined functionally, following

previous work in typology, especially Haspelmath (2017) and Koptjevskaja-Tamm (2003):

(21) Possessive construction
A possessive construction is a construction that expresses ownership (e.g., *my name*), kinship (e.g., *my mother*), or part-whole relationship (e.g., *my head*).

As is well-known, there are languages that distinguish alienable possession and inalienable possession (Bugaeva et al. 2022; Haspelmath 2017; Nichols 1988). Some languages use the juxtaposition strategy for inalienable possession. For example, Kakabe (kke; Mande, Western Mande) and Wappo (wao; Yuki-Wappo, Wappo) use the juxtaposition strategy only for inalienable possession, as in (22) and (23).

(22) Kakabe (Mande, Western Mande)

a. Alienable possession (Vydrina 2017: 92)
mùséé *là* *sáákòè*
woman.ART POSS bag.ART
'woman's bag'

b. Inalienable possession (Vydrina 2017: 92)
mùséè *bólè*
woman.ART hand.ART
'woman's hand'

(23) Wappo (Yuki-Wappo, Wappo)

a. Alienable possession (Thompson et al. 2006: 26)
ah *te-me?* *papelˈ* *peh-khi?*
1SG.NOM 3SG-GEN book look-STAT
'I am looking at his/her book.'

b. Inalienable possession (Thompson et al. 2006: 15)
c'ic'a *khap-i* *ke?te-khi?*
bird wing-NOM broken-STAT
'The bird's wing is broken.'

Interestingly, in Apurinã (apu; Arawakan, Southern Maipuran), inalienable nouns use the noun juxtaposition strategy, and nouns require an unpossession marker when unpossessed, as illustrated in (24).

(24) Apurinã (Arawakan, Southern Maipuran)

a. Inalienable possession (Facundes 2000: 152)
kema *kuwu*
tapir head
'tapir's head'

b. Unpossession (Facundes 2000: 153)
kuwĩ-txi
head-NPOSS
'the head'

However, there are also languages that use the juxtaposition strategy only for alienable possession. This is the case with Ndjébbana (djj; Maningrida, Nakkara-Ndjebbana) in (25).

(25) Ndjébbana (Maningrida, Nakkara-Ndjebbana)

a. Alienable possession (McKay 2000: 195)
marddúrddiba *ngáyabba*
heart I
'my heart'

b. Inalienable possession (McKay 2000: 208)
díla-ngaya
eye-her
'her eye'

In this paper, all cases of noun juxtaposition are taken into account regardless of the type of possession, namely alienable or inalienable. The reason for this is that the distinction between alienable and inalienable depends on how the terms are defined. Previous studies show disagreement in the usage of the terminology *alienability*. In Cristofaro (2023), the terms *alienable* and *inalienable* are defined from a functional

(notional) perspective, whereas in Nichols (1988), they are defined from a formal (hybrid) perspective. The functional definition classifies nouns as (in)alienable based on their inherent meaning, such as kinship terms and body parts, and these classifications remain consistent across languages. In contrast, the formal (or hybrid, in the sense of Haspelmath 2024a) definition identifies nouns as inalienable when they use a shorter (or zero) form in the alienability split. Consequently, the nouns classified as inalienable vary from language to language. Since the present study focuses on the syntactic structure (strategy), specifically noun juxtaposition, and investigates whether it exhibits ambiguity in relation to other functions, rather than within a single function, I do not explore which subfunctions of possession are typically expressed by noun juxtaposition.

3.4. Conjunction

Nouns in juxtaposition can express a conjunctive relationship. This is exemplified in Southern Pomo (peq; Pomoan, Russian River and East), Bukiyip (ape; Nuclear Torricelli, Kombio-Arapesh-Urat), and Matses (mcf; Pano-Tacanan, Panoan) as in (26).

(26) a. Southern Pomo (Pomoan, Russian River and East; Walker 2020: 335)
miy:a-me-Ø *miy:a-t^{h}e-Ø*
3-father-AGT 3-mother-AGT
'her father and mother'

b. Bukiyip (Nuclear Torricelli, Kombio-Arapesh-Urat; Conrad & Wogiga 1991: 64)
ot-uk *élmatok* *at-unú* *élman*
one-3SG.F woman one-3SG.M man
'one woman and one man'

c. Matses (Pano-Tacanan, Panoan; Fleck 2003: 805)
senta-n *chëshëid-n*
uakari.monkey-ERG spider.monkey-ERG
'uakari monkeys and spider monkeys'

As mentioned earlier, I include conjunction because it involves NPs, and there is potential

ambiguity between predication, adnominal possession, and conjunction. However, it should be noted that the examples in (26) may deviate from the definition of noun juxtaposition in (8), according to Haiman (1983).[6] This author argues that iconic markers also function as coordination markers. For example, in (26c), ergative markers are used not only as ergative markers but also as coordination markers. However, I do not follow this idea. I have two reasons for this. First, dedicated coordination markers can be used regardless of the presence of these iconic markers. As shown in (27a), the coordination marker *chedo* can be used when iconic markers are present, whereas the juxtaposition strategy can also be employed without these iconic markers, as in (27b). Thus, the difference between the presence and absence of iconic markers does not contribute to the meaning of coordination.

(27) Matses (Pano-Tacanan, Panoan; Fleck 2003: 803; 812)

a. *mëcueste-n* *capa* *chedo-n*
agouti-ERG squirrel too-ERG
'agoutis and squirrels'

b. *titado* *pachid*
peach.palm manioc
'peach plam fruits and/or manioc'

Second, iconic markers can be found in contexts other than coordination. In (28a), iconic markers are used in predication, and in (28b), they are used in adnominal possession.

(28) a. Russian (Indo-European, Balto-Slavic)
Moj *otec* *moj* *učitel'*.
1SG.POSS.M father 1SG.POSS.M teacher
'My father is my teacher.'

b. Tima (tms; Katla-Tima, Tima; Alamin Mubarak 2009: 131)
k-ʊbɑy *k-ahʊnɛn*
SG-cup SG-woman
'woman's cup'

[6] I owe this point to a reviewer.

In coordination, by definition, units of the same syntactic status are construed together. Since they share the same status, they tend to have iconic markers. However, this does not mean that these iconic markers function as coordination markers.

This paper deals only with conjunction, a type of coordination. Phrase coordination is typically subdivided into conjunction and disjunction based on function, and noun juxtaposition is sometimes used for disjunction as well, as in (29).

(29) Ngarinyin (ung; Worrorran, Ngarinyin; Spronck 2015: 38)
kanangkurr *aru* *dolad* *warndij* *mo$_2$-y$_2$i-nyi-nu*
dog snake hole create 3N.SBJ-be-PST-2SG.OBJ
'It could become a dog, snake, or hole for you.'

However, this paper does not consider disjunction because information on disjunction in reference grammars tends to be much less detailed than conjunction. *Conjunction* is defined as follows:

(30) Conjunction (cf. Croft 2022: 680; 682; Haspelmath 2007b: 1)
Conjunction is a type of coordination that is a syntactic construction in which two or more units of the same status are construed into a larger unit and represents some sort of grouping together in the relevant context.

For conjunction, some languages allow the connection of more than two conjuncts.[7] When coordinating more than two coordinands (multiple coordinands), many languages permit the omission of function indicators. This is illustrated in the following example (31) from Haspelmath (2007b: 12).

(31) West Greenlandic (Eskimo-Aleut, Eskimo; Haspelmath 2007b: 12 from Fortescue 1984: 127)
tulu-it *qallunaa-t* *kalaall-il=lu*
Englishman-PL Dane-PL Greenlander-PL=and
'Englishmen, Danes and Greenlanders'

The first two coordinands in (31) and their English translation do not have any marker.

[7] The use of the terms is based on Croft (2022); Haspelmath (2004; 2007b).

In this sense, this falls under the definition of noun juxtaposition in the present paper. However, I do not consider such examples because the function indicator (*lu* in (29)) appears to reflect the relationship of the entire phrase.

The definition of conjunction above excludes the construction that is called *inclusive constructions* in Goddard (1985: 51) and Langlois (2004: 118–19), as well as *associative constructions* in Dunn (1999: 172). This is illustrated in the following examples (32) from Chukchi (ckt; Chukotko-Kamchatkan, Chukotian) and Pitjantjatjara (pjt; Pama-Nyungan, Desert Nyungic).

(32) a. Chukchi (Chukotko-Kamchatkan, Chukotian; Dunn 1999: 172)
ətləyə-t *əmmemə*
parent-3PL.ABS mother.3SG.ABS
'the father and mother'

b. Pitjantjatjara (Pama-Nyungan, Desert Nyungic; Langlois 2004: 118)
Annie-nya *tjana* *Sydney-lakutu* *a-nu.*
Annie-NOM 3PL.NOM Sydney-ALL go-PST
'Annie and her friends went to Sydney.'

In these examples, the reference of one of the coordinands (*əmmemə* and *Annie-nya*, respectively) is included in the other coordinand (*ətləyə-t* and *tjana*, respectively). In this sense, these coordinands do not have the same status.

The definition of noun juxtaposition in (8) excludes examples which contain function indicators from the scope of the survey. For instance, Telugu (tel; Dravidian, South Dravidian) and Sanuma (xsu; Yanomamic, Sanumá) examples in (33) and (34) are not classified as noun juxtaposition, because lengthened final vowels can be considered indicators in Telugu, and a summary phrase can be considered an indicator in Sanuma, respectively.

(33) Telugu (Dravidian, South Dravidian; Krishnamurti & Gwynn 1985: 325)

a.	*aayana*	b.	*miiru*	c.	*aayanaa miiruu*
	he		you.PL		he and you

(34) Sanuma (Yanomamic, Sanumá; Borgman 1990: 35)

pumotomö	*a,*	*samonamaniwö*	*a,*	*ĩ*	*naha*	*kule-i,*	*tökö*	
opossum.man	3SG	bee.man	3SG	REL	like	be-INDF	3DU	

ku-kö-ma
stay-FOC-COMPL
'Opossum-man and Bee-man stayed.'

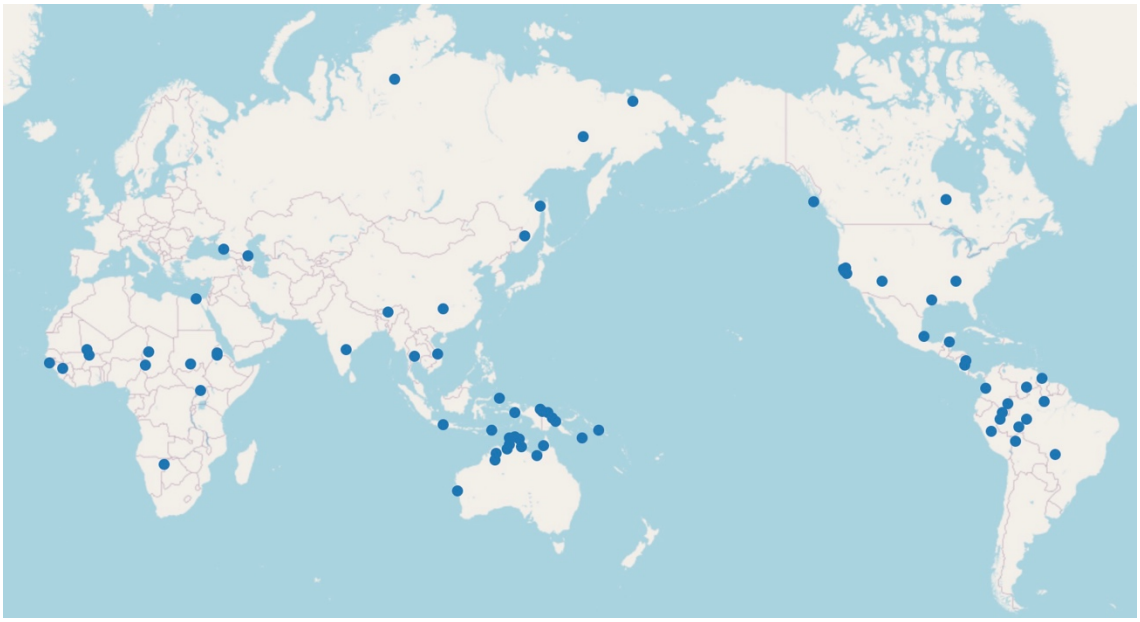
4. Language sample

Several sampling methods have been proposed in the typological literature (Miestamo et al. 2016; Rijkhoff et al. 1993; Di Garbo & Napoleão de Souza 2023; Rijkhoff & Bakker 1998, among others). Since every sampling method has its own strengths, the type of typological sampling best suited depends on the research question. Probability sampling, for example, is used to examine correlations and tendencies, while variety sampling is used for exploratory research, specifically, for examining variation.

Insofar as the present study aims to investigate whether ambiguity plays a more important role than efficiency in explaining the use of certain linguistic forms, a probability sample seems more appropriate. However, another aim of this paper, such as investigating which functions are typically expressed by noun juxtaposition cross-linguistically, requires a variety sample. Therefore, independence and representativeness are equally important for this study. To ensure the independence of languages, the sample includes one and only one language from each genus as proposed by Miestamo et al. (2016). Even though in their method the areal stratification is made at the level of macro-areas and the number of languages in each macro-area is proportional to the number of genera within that macro-area, this study does not follow that approach. The reason for this is that this study also aims to examine areality in relation to the use of noun juxtaposition. As mentioned in Section 2, since the hypothesis in Frajzyngier et al. (2002) is proposed based on the sample biased toward African languages, there is a possibility that the use of noun juxtaposition exhibits some areal patterns. Thus, this paper has an equal number of languages per macro-area.

In this way, I survey a sample of 72 languages, consisting of 12 languages from each macro-area, as shown in Map 1. The decision regarding the number of languages in the sample is somewhat arbitrary, but 12 languages seem sufficient to investigate areality, because Frajzyngier et al. (2002) include 11 African languages. As defined

in (8), noun juxtaposition is a structure that serves as one of the strategies for predication, possession, and conjunction. Thus, not all languages use it for these three functions. For example, I could not find noun juxtaposition used for these three functions in Molalla (mbe; Isolate, North America; Pharris 2006), Choguita Rarámuri (tar; Uto-Aztecan, Southern Uto-Aztecan; Caballero 2022), and Karelian (krl; Uralic, Finnic; Novak et al. 2022). The sample intentionally excludes languages where noun juxtaposition is not used for the three functions in question. Almost all sources are reference grammars or grammar sketches. The selection of languages is based on data accessibility. Complete information on the sample and sources is provided in Appendix A. All information on the geographical and genealogical distribution of languages is based on Glottolog 5.0.



Map 1. Languages of the sample.[8]

5. Results of the worldwide survey

In this section, I present the results of the survey. Since this study investigates a one-form-three-function relationship, there are seven logically possible language types, as shown in (35).

[8] All maps in this paper were created with the help of Lingtypology (Moroz 2017).

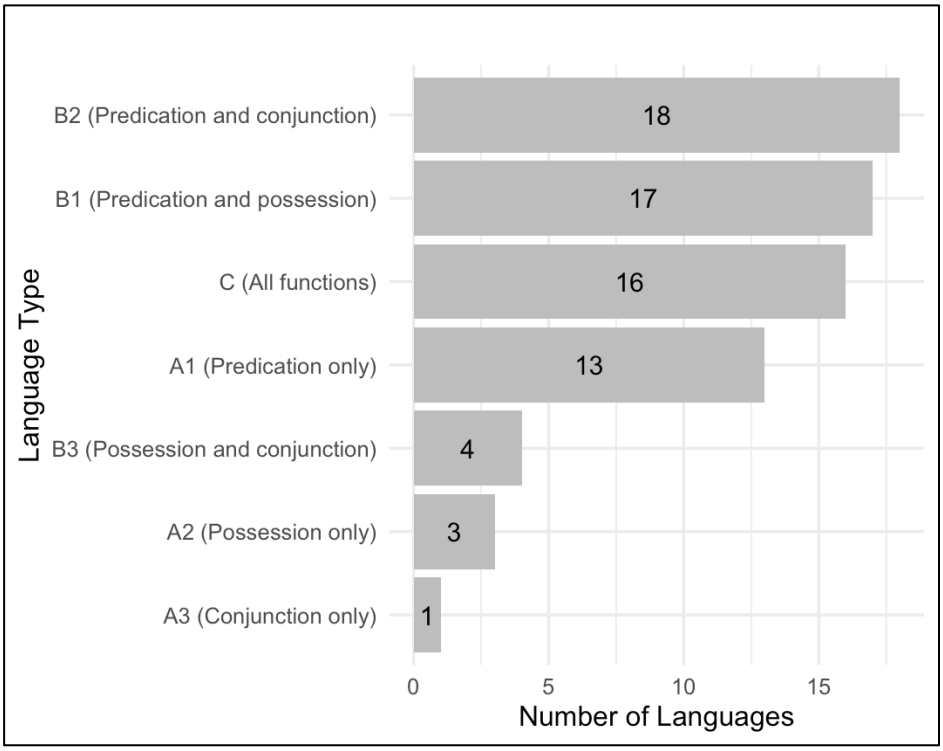
(35) Language types based on the distribution of noun juxtaposition

a. (one function) Predication only (A1)
Possession only (A2)
Conjunction only (A3)

b. (two functions) Predication and possession (B1)
Predication and conjunction (B2)
Possession and conjunction (B3)

c. (all functions) Predication, possession, and conjunction (C)

All types are attested in the sample, but the ratio is not equal. For example, the predication and conjunction type (B2) accounts for 25%, while the conjunction only type (A3) accounts for just 1.4% (see Graph 1).



Graph 1. Distribution of noun juxtaposition in the sample.

The remainder of this subsection provides concrete examples for each language type.

5.1. Predication only type (A1)

There are thirteen languages in this type in the sample. An example is Dazaga (dzg;

Saharan, Western Saharan), where noun juxtaposition can be used for predication, as illustrated in (36a), but overt markers are required to express possession and conjunction, as in (36b) and (36c).

(36) Dazaga (Saharan, Western Saharan; Walters 2016: 145; 128; 173)

a. Predication

àrìí *áì* *ájá* *nír*
woman this mother 1SG.POSS
'This woman is my mother.'

b. Possession

fúrcì *gʷɔ̀ní=ŋà*
dung camel=GEN.SG
'camel's dung'

c. Conjunction

fírí-a=jɛ́ *képtí=jɛ̀*
arrow-PL=and bow=and
'arrows and a bow'

5.2. Possession only type (A2)

Three languages in the sample fall into this type. In Tommo So, a copula and associative markers are used for predication and conjunction, respectively, while possession can be expressed through noun juxtaposition, as shown in (37).

(37) Tommo So (Dogon, Escarpment Dogon; McPherson 2013: 340; 190; 211)

a. Predication

Ú *mí* *ánìgè=ɲì*
2SG 1SG friend=COP
'You are my friend.'

b. Possession

Sáná *bàbè*
Sana uncle
'Sana's uncle'

c. Conjunction
ɛ̀ŋjɛ́ = le *jàmdúlu = le*
chicken = ASSOC donkey = ASSOC
'a chicken and a donkey'

5.3. Conjunction only type (A3)

Only one language, specifically Patwin (pwi; Wintuan, Patwin) is classified under this type in the sample. In this language, noun juxtaposition can express conjunction, as in (38c), while predication and possession require function indicators, as in (38a) and (38b).

(38) Patwin (Wintuan, Patwin; Lawyer 2015: 294; 93; 190)
a. Predication
ʔew *ʔi-s* *bi·t*
PROX.SG.SBJ COP-FIN meadowlark
'That is a meadowlark.'

b. Possession
wita-no *nun*
man-POSS gun
'the man's gun'

c. Conjunction
san-č'iyak *kat^{h}it-se·ktu*
sun-old.man falcon-chief
'Old Man Sun and Falcon Chief'

5.4. Predication and possession type (B1)

There are seventeen languages in this type. For instance, Labwor can use noun juxtaposition for both predication and possession, but it requires an overt marker for conjunction, as illustrated in (39).

(39) Labwor (Nilotic, Western Nilotic; Heine & König 2010: 30; 61; 98)

a. Predication
mánón *bɔ̀ɔ́*
that bɔɔ
'It is bɔɔ vegetable.'

b. Possession
ɔ̀t *dhákɔ́*
house woman
'woman's house'

c. Conjunction
ɛ̀cúɔ̀ *gín_kí* *dhákɔ́*
man and woman
'the man and the woman'

5.5. Predication and conjunction type (B2)

Eighteen languages in the sample fall into this type. In Nhanda (nha; Pama-Nyungan, South-West Pama-Nyungan), noun juxtaposition can express both predication and conjunction, but possession requires a genitive marker, as illustrated in (40).

(40) Nhanda (Pama-Nyungan, South-West Pama-Nyungan; Blevins 2001: 66; 57; 134)

a. Predication
ngana-bagaa *inya* *uthu?*
who-PROP this dog
'Whose dog is this?'

b. Possession
uthu-wu *thudu-ra*
dog-GEN meat-3OBL
'the dog's meat'

c. Conjunction
acijadi-wana *mirla-wana*
clothes-1PL rug-1PL
'our clothes and our rugs'

5.6. Possession and conjunction type (B3)

Four languages in the sample are classified under this type. In Matses, noun juxtaposition can be used for both possession and conjunction, but a copula is required for predication, as shown in (41).

(41) Matses (Pano-Tacanan, Panoan; Fleck 2003: 944; 764; 805)
a. Predication
ubi *dësi* *ne-e-c*
1ABS Dësi COP-NPST-IND
'I am Dësi.'

b. Possession
bucu *podo*
cecropia leaf
'leaves of cecropia trees'

c. Conjunction
senta-n *chëshëid-n*
uakari.monkey-ERG spider.monkey-ERG
'Uakari monkeys and spider monkeys'

5.7. All functions type (C)

Sixteen languages in the sample can use noun juxtaposition for all functions, as illustrated in (42).

(42) Ndjébbana (Maningrida, Nakkara-Ndjebbana; McKay 2000: 292; 195; 306)
a. Predication
Nǰanabbárdakka *yírrìddjanga.*
trevally (fish) Yírrìddjanga
'The trevally is Yírrìddjanga.'

b. Possession
marddúrddìba *ngáyabba*
heart I
'my heart'

c. Conjunction
warakkála, *karndóya*
long.yam round.yam
'long yams and round yams'

6. Discussion

In this section, I observe the distributional tendencies of noun juxtaposition and make several generalizations based on the results presented in the previous section. In addition, I make a theoretical suggestion based on these observations. All data on the distribution are presented in Appendix B.[9]

6.1. Distributional tendencies of noun juxtaposition

As shown in Graph 1 in Section 5, noun juxtaposition can express predication in 64 languages (89%) of the sample. One observation can be made at this point.

(43) Observation 1
There is a strong tendency for noun juxtaposition to be used for predication.

Also, in many cases, noun juxtaposition can be used for two or three functions. In the sample, 55 languages (76%) exhibit this tendency.

(44) Observation 2
Many languages use noun juxtaposition for more than one function.

Since these two observations represent strong tendencies, the cases where they do not apply deserve some attention, specifically A2 (Possession only), and A3 (Conjunction

[9] Almost all of the examples of noun juxtaposition considered in the present paper can be found in CrossGram (https://crossgram.clld.org/).

only). In the remainder of this subsection, I examine each type in detail.

Concerning A2, all the languages classified under A2 in the sample are African languages, and at this point, an observation can be formulated as in (45). This type may be commonly observed in African languages, which, along with the African pattern below, may explain why Frajzyngier et al. (2002) reached their conclusion: noun juxtaposition can be used for either predication or possession within a single language (see Section 2). It should be noted, however, that this observation cannot be generalized to all languages classified as A2 being African because I am aware of non-African languages that can also be classified under this type outside of the sample, such as Welsh (cym; Indo-European, Celtic; see Koptjevskaja-Tamm 2002: 144; 2003: 649).

(45) Observation 3
A language that uses noun juxtaposition for only possession among predication, possession, and conjunction is an African language.

Languages that use noun juxtaposition only for conjunction (A3) are very rare. Only one language, Patwin, belongs to this type in the sample. This rarity seems to allow for a generalization like (46) as a very strong tendency, but I am skeptical about such a generalization.

(46) Generalization (tentative)
If a language uses noun juxtaposition for conjunction, the language uses at least one of the other functions.

I present three reasons for caution. First, the data are not sufficient to support such a generalization. As noted in Observation 1, most of the world's languages that use noun juxtaposition for at least one of the three functions in question can use it for predication. Only eight languages do not use it for predication in the sample.

Second, it seems that there is no correlation between the use of noun juxtaposition for conjunction and that for predication or possession. As is well-known, many conjunction markers have been grammaticalized or borrowed recently (Haspelmath 2007b: 8; Mithun 1988). Consequently, noun juxtaposition for conjunction tends to be marginalized into specific functions or is altogether replaced by other marking strategies (Stassen 2000: 10). As Mithun (1988: 351) notes, the way markers emerge

varies from language to language, but noun juxtaposition has been replaced with a non-juxtaposition strategy as the general trend all over the world (Stassen 2000: 10). Thus, the result of the present survey may be considered a reflection of this process of replacement that has been completed, is currently ongoing or is expected to occur in the future.[10] Indeed, the emergence of a non-juxtaposition strategy is often attributed to ambiguity in the interpretation of noun juxtaposition. For example, Borise & É Kiss (2023) argue that conjunction markers have emerged in Khanty (Uralic, Khantyic) due to ambiguity. While this explanation may hold true, the results of the present study do not support the idea of an ambiguity between the use of noun juxtaposition for conjunction and its use for predication or possession. This is because almost all languages that use noun juxtaposition for conjunction also make use of it for one or two other functions. If a language developed conjunction markers to avoid ambiguity among the three functions, we would expect to find more languages classified as A3, since noun juxtaposition would not exhibit ambiguity if it were solely dedicated to conjunction. Thus, while ambiguity might arise in the interpretation of subfunctions within a function (e.g., among conjunction, disjunction, and adversative coordination within coordination), such ambiguity does not seem to exist among different functions.

Third, the scope of each function is different. The present study addresses three functions: predication, possession, and conjunction. The scope of conjunction is smaller than that of predication and possession. As mentioned in Section 3, both predication and possession have several subfunctions, which are all taken into consideration. In contrast, conjunction, as considered in this study, is a subfunction of coordination. Thus, the potential for noun juxtaposition to be used for conjunction is likely lower than that for predication and possession.

Thus, languages rarely use noun juxtaposition exclusively for conjunction, however, this fact may not be generalized as in (46).

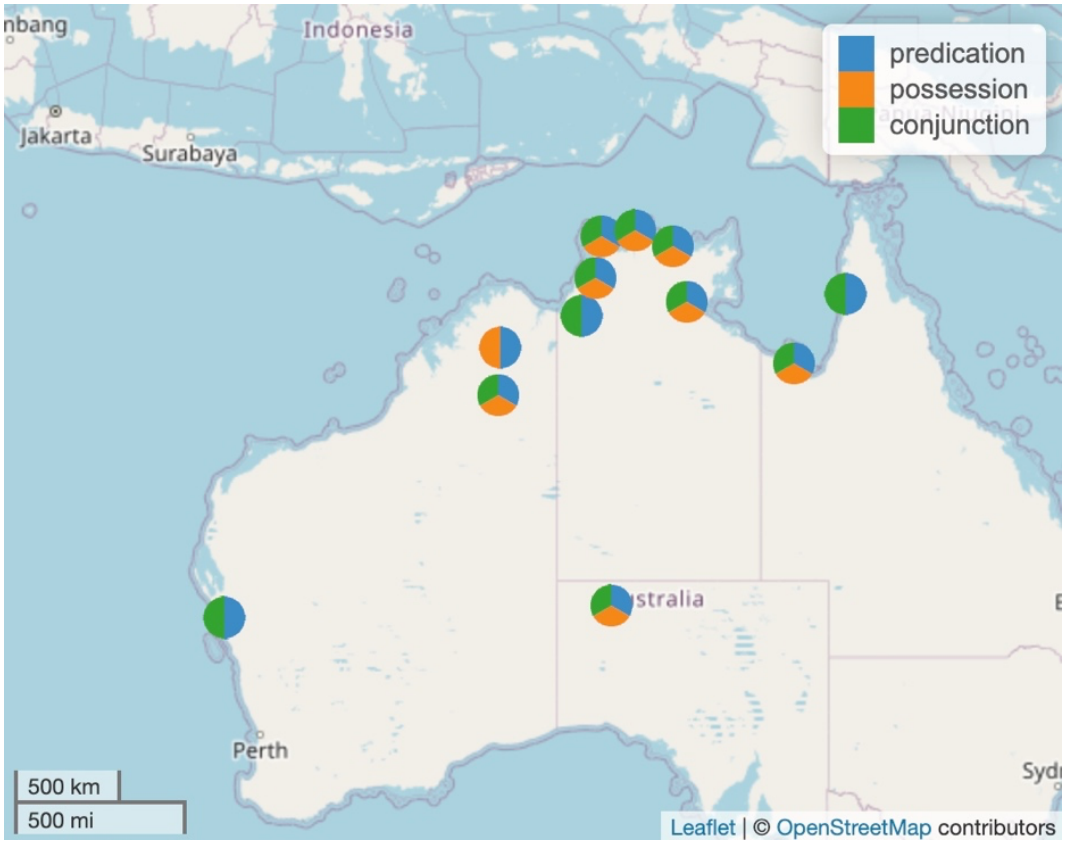
6.2. Areal patterns

In the previous subsection, I presented general observations based on the results. In this subsection, I report on several areal patterns.

[10] This fact is also concretely illustrated in relatively recent grammars, such as those of Papuan Malay (pmy; Austronesian, Malayo-Polynesian; Kluge 2017: 558) and Sumerian (sux; Isolate, Eurasia; Jagersma 2010: 95–100).

6.2.1. Australia

Let us first consider the Australian languages.



Map 2. Distribution of noun juxtaposition in Australian languages.[11]

As shown in Map 2, all Australian languages in the sample use noun juxtaposition for two or three functions, one of which is predication. In other words, types not involved in predication, such as the possession and conjunction type (B3) are absent in Australia.

(47) Australian pattern
Australian languages typically use noun juxtaposition for two or three functions, one of which is predication.

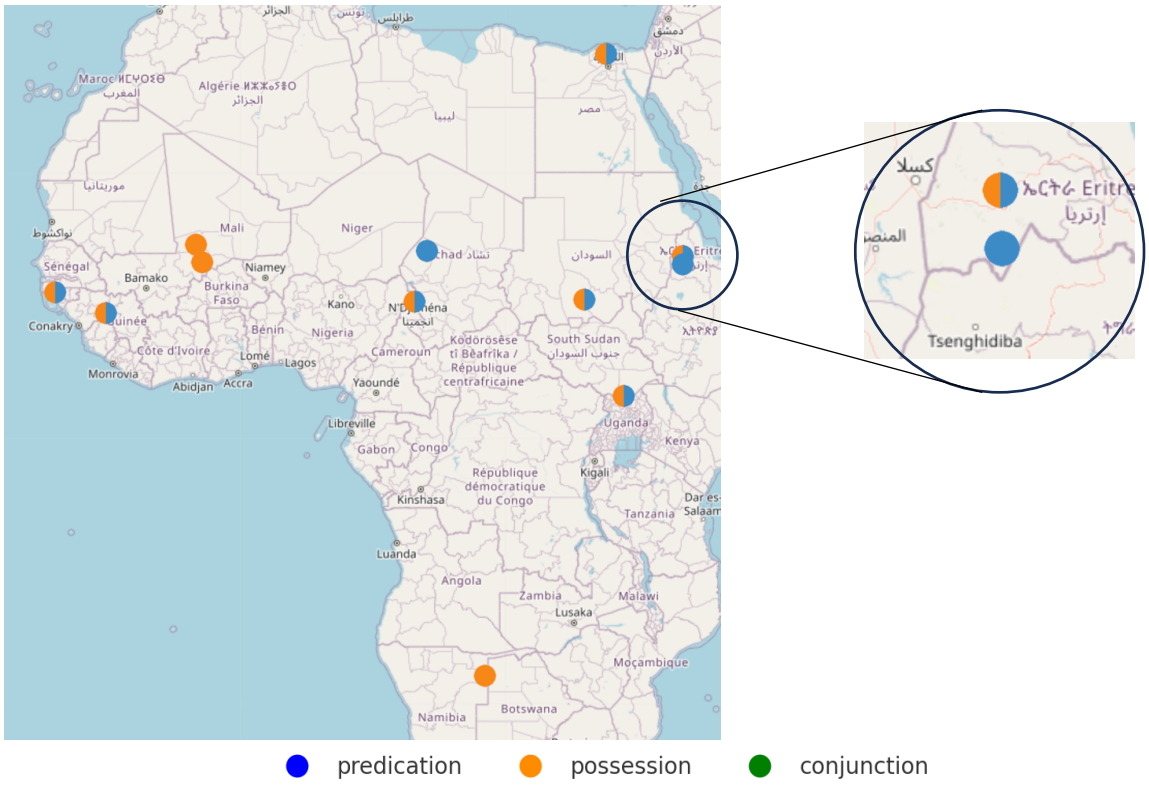
The extensive use of noun juxtaposition in Australian languages is well-known (e.g.,

[11] In all the maps below, blue indicates the use of noun juxtaposition for predication, orange indicates possession, and green indicates conjunction.

Evans 1995: 313; Sadler & Nordlinger 2010). However, (47) elaborates on this by providing empirical information about functions expressed by noun juxtaposition.

6.2.2. Africa

As mentioned in Observation 3, languages classified as the possession only type (A2) are typically found among African languages. This can be considered as an African characteristic. However, African languages exhibit another pattern as well.



Map 3. Distribution of noun juxtaposition in African languages.

As shown in Map 3, no African language in the sample uses noun juxtaposition for conjunction.

(48) African pattern
African languages use noun juxtaposition for predication and/or possession.

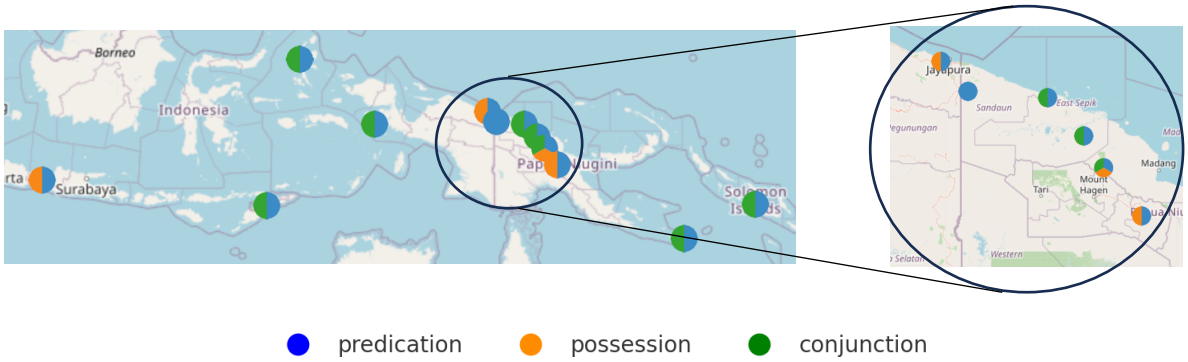
Regarding the absence of the juxtaposition strategy for conjunction in African languages, this has already been noted by Stassen (2000: 9). Since the results of the present survey replicate his findings, this can be generalized as follows:

(49) Generalization 1: Noun juxtaposition in African languages
African languages rarely use noun juxtaposition for conjunction.

As mentioned earlier, the fact that Frajzyngier et al. (2002) focus only on predication and possession (even though noun juxtaposition can also be used for conjunction) may be attributable to this African pattern.

6.2.3. Papunesia

Papunesian languages also show an interesting pattern, as illustrated in Map 4.



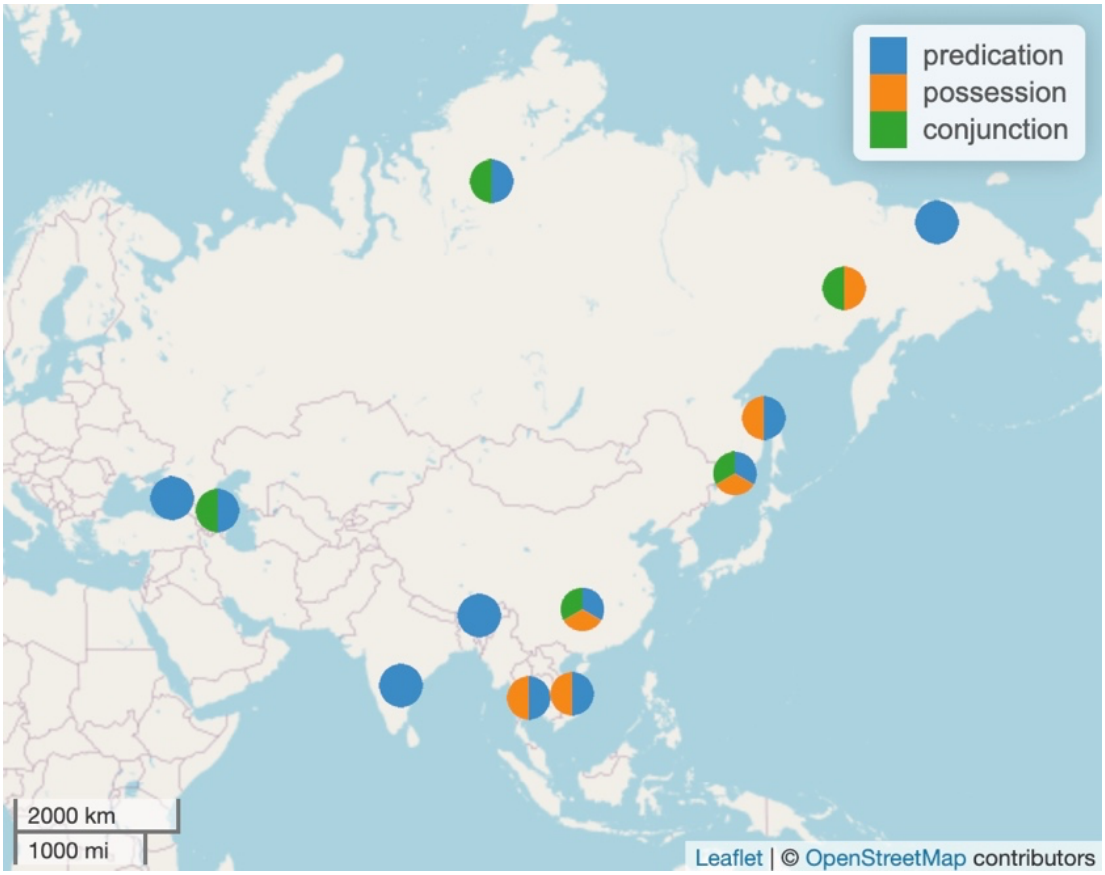
Map 4. Distribution of noun juxtaposition in Papunesian languages.

As shown in Map 4, all Papunesian languages in the sample use noun juxtaposition for predication.

(50) Papunesian pattern
Papunesian languages use noun juxtaposition at least for predication.

6.2.4. Eurasia

The distribution of language types among Eurasian languages is illustrated in Map 5.



Map 5. Distribution of noun juxtaposition in Eurasian languages.

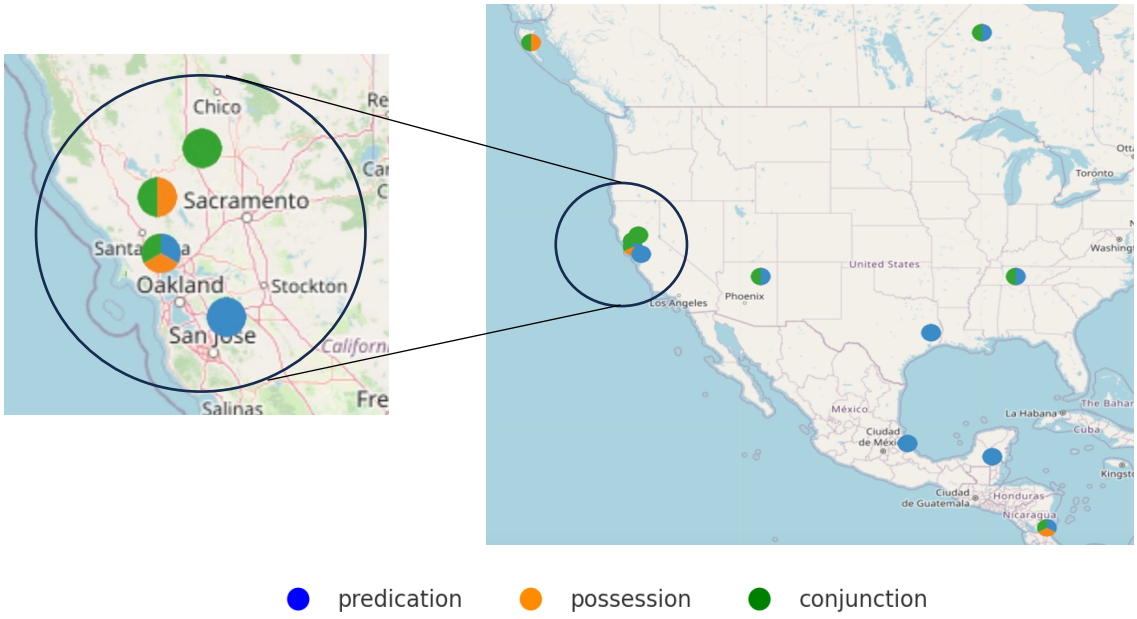
Many Eurasian languages in the sample use noun juxtaposition for more than one function. When it is used only for one function, it is for predication. Interestingly, possession seems to appear in eastern languages on the map. However, further research is required to determine whether the use of noun juxtaposition for possession is a characteristic feature of eastern languages in Eurasia.[12]

(51) Eurasian pattern
Eurasian languages do not use noun juxtaposition exclusively for possession or conjunction.

6.2.5. North America

The distribution of language types among North American languages is illustrated in Map 6.

[12] I owe this point to the editors.



Map 6. Distribution of noun juxtaposition in North American languages.

All types except for A2 are attested among North American languages.

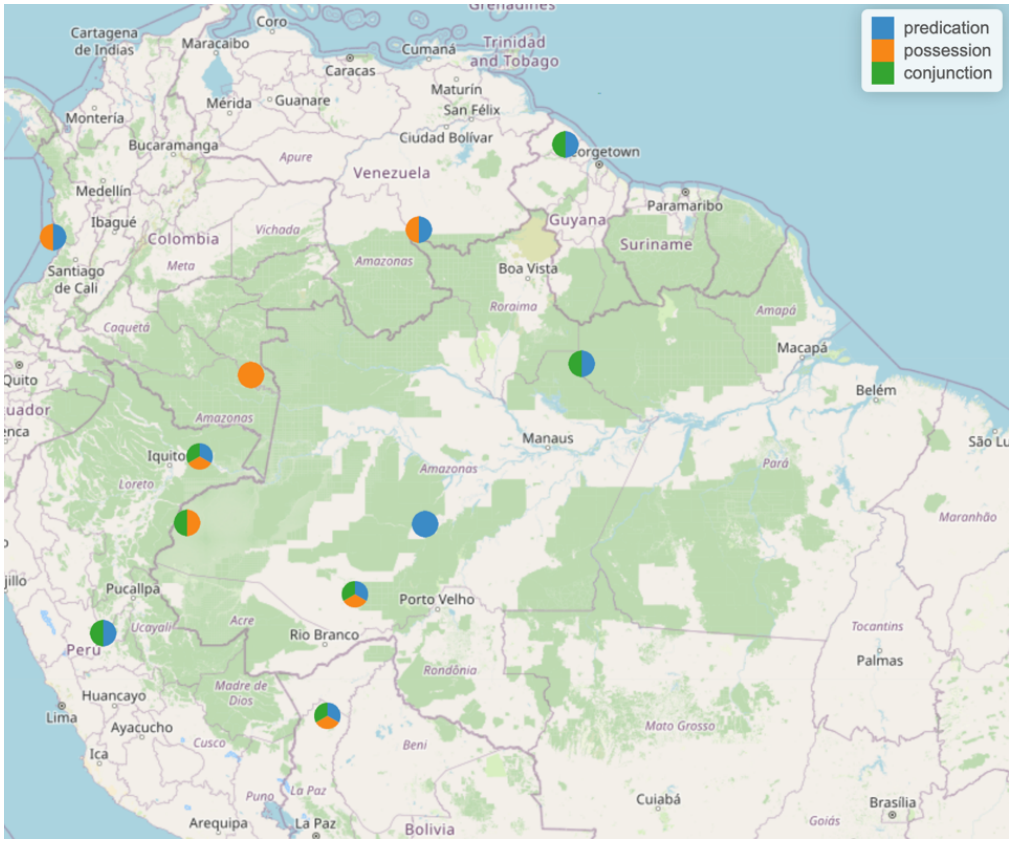
(52) North America pattern
Nouth American languages do not use noun juxtaposition exclusively for possession.

6.2.6. South America

The distribution of language types among South American languages, as illustrated in Map 7, is quite similar to that of Eurasian languages.

Many South American languages in the sample use noun juxtaposition for more than one function. When it is used only for one function, it is for predication.

(53) South American pattern
South American languages do not use noun juxtaposition exclusively for possession or conjunction.



Map 7. Distribution of noun juxtaposition in South American languages.

6.3 Ambiguity versus efficiency

In the preceding two subsections, we observed the results of the present survey concerning generality and areality. Since the investigation in terms of areality does not contradict general observations in (43) and (44), these two observations seem to be generalizable. Therefore, the following generalizations can be made:

(54) Generalization 2: Functions typically expressed by noun juxtaposition
If a language can use a structure of noun juxtaposition, it is predominantly used for predication in most cases.

(55) Generalization 3: The number of functions expressed by noun juxtaposition
If a language can use a structure of noun juxtaposition, it typically serves more than one function.

Generalization 3 contradicts the observation proposed by Frajzyngier et al. (2002),

which challenges the hypothesis that the use of noun juxtaposition is constrained by ambiguity avoidance. Rather, the present study suggests that the use of noun juxtaposition should be explained by efficiency (Hawkins 2014: Section 2.2; Haspelmath 2017). Since noun juxtaposition, by definition, can be considered the most efficient form in terms of formal length, it is potentially the most ambiguous. Ambiguity and efficiency are important factors in explaining the use of specific forms (e.g., Hankamer 1973; Levshina 2022, respectively), though they can oppose each other. When speakers use simpler forms, listeners may misunderstand the speaker's intentions. Therefore, the question of which is prioritized – ambiguity or efficiency – has been a topic of discussion in explaining the use of certain forms. The present study implies that ambiguity across functions does not significantly influence the use of certain forms, and it is not always avoided, as demonstrated through the examination of noun juxtaposition, a structure well-suited for investigating this issue. Instead, human languages tend to prefer simpler (more efficient) forms, with ambiguity being resolved through other means, such as context, word order, and/or phonological factors.[13] Thus, the use of a certain linguistic form (noun juxtaposition in this case) should be explained in terms of efficiency rather than ambiguity avoidance. This is consistent with the claims made by Piantadosi et al. (2012) and Wasow (2015).

7. Conclusion

In this paper, I have investigated noun juxtaposition, using a balanced sample of 72 languages, and claimed that the use of certain linguistic forms, such as noun juxtaposition should be explained by efficiency rather than ambiguity. Although noun juxtaposition is used worldwide, it has rarely been studied cross-linguistically. One notable exception is the work of Frajzyngier et al. (2002), who argue that the use of noun juxtaposition is constrained by ambiguity avoidance. However, the present paper does not support this hypothesis and finds that their observations are biased toward African areal patterns. Rather, this study finds that languages predominantly use noun juxtaposition for predication, and it typically serves more than one function. Since noun juxtaposition is, by definition, the most efficient yet ambiguous form, these generalizations suggest that efficiency is more prioritized over ambiguity in explaining the use of noun juxtaposition. Also, ambiguity across functions does not

[13] The investigation of the factors that contribute to resolving ambiguity falls outside the scope of the present study and requires further research.

significantly influence the use of certain forms, and it is not always avoided. This is consistent with the claims made by Piantadosi et al. (2012) and Wasow (2015). Thus, the present paper can be regarded as a case study that contributes to the discussion of whether ambiguity or efficiency is prioritized in language.

Acknowledgements

I am deeply grateful to my advisor, Martin Haspelmath, for the useful comments and advice. I also thank Chihiro Taguchi for his assistance with stylistic improvements. In addition, I thank two anonymous reviewers and the editors (particularly, Francesca Di Garbo). This work was partly supported by JST SPRING, grant number JPMJSP2110.

Abbreviations

1 = 1st person
2 = 2nd person
3 = 3rd person
ABS = absolutive
AGT = agentive
ALL = allative
ART = article
ASSOC = associative
COM = comitative
COMPL = completive
COP = copula
DIST = distal
DU = dual
ERG = ergative
F = feminine
FIN = finite
FOC = focus
GEN = genitive
IND = indicative
INDF = indefinite
LOC = locative
M = masculine
N = neuter
N- = non-
NC = noun class
NEG = negative
NOM = nominative
OBJ = object
OBL = oblique
PL = plural
POSS = possessive
PRED = predicative
PROP = proprietive
PROX = proximate
PST = past
REL = relative
SBJ = subject
SG = singular
STAT = stative aspect

References

Alamin Mubarak, Susan. 2009. *Tima Word Structure: Noun and Verb*. Khartoum: University of Khartoum PhD dissertation.

Barlow, Russell. 2023. *A grammar of Ulwa (Papua New Guinea)*. Berlin: Language Science Press.

Bender, Marvin Lionel. 1996. *Kunama*. Vol. 59. Munich: Lincom Europa.

Birk, David. 1976. *The Malakmalak Language, Daly River (Western Arnhem Land)*. Vol. 45. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Blevins, Juliette. 2001. *Nhanda: An Aboriginal Language of Western Australia* (Oceanic Linguistics Special Publication). Vol. 30. Honolulu: University of Hawaii Press.

Bodt, Timotheus Adrianus. 2020. *Grammar of Duhumbi (Chugpa)*. Leiden: Brill.

Bolles, David & Bolles, Alejandra. 2014. *A grammar and anthology of the Yucatecan Mayan language*. Milford, CT: Ms.

Borgman, Donald M. 1990. Sanuma. In Derbyshire, Desmond C. & Pullum, Geoffrey K. (eds.), *Handbook of Amazonian Languages 2*, 15–248. Berlin: Mouton de Gruyter.

Borise, Lena & É Kiss, Katalin. 2023. The emergence of conjunctions and phrasal coordination in Khanty. *Journal of Historical Linguistics* 13(2). 173–219. https://doi.org/10.1075/jhl.21016.kis.

Bowe, Heather J. 1990. *Categories, constituents and constituent order in Pitjantjatjara: an Aboriginal language of Australia* (Theoretical Linguistics). London, New York: Routledge.

Bugaeva, Anna, Johanna Nichols & Bickel, Balthasar. 2022. Appositive possession in Ainu and around the Pacific. *Linguistic Typology* 26(1). 43–88. https://doi.org/doi:10.1515/lingty-2021-2079.

Caballero, Gabriela. 2022. *A grammar of Choguita Rarámuri: In collaboration with Luz Elena León Ramírez, Sebastián Fuentes Holguín, Bertha Fuentes Loya and other Choguita Rarámuri language experts* (Comprehensive Grammar Library). Berlin: Language Science Press.

Capell, Arthur. 1962. *Some linguistic types in Australia*. Sydney: University of Sydney.

Chapman, Shirley & Derbyshire, Desmond C. 1991. Paumari. In Derbyshire, Desmond C. & Pullum, Geoffrey K. (eds.), *Handbook of Amazonian Languages 3*, 161–352. Berlin: Mouton de Gruyter.

Coate, Howard H. J. & Oates, Lynette Frances. 1970. *A Grammar of Ngarinjin, Western Australia*. Canberra: Australian Institute of Aboriginal Studies.

Conrad, Robert J. & Wogiga, Kepas. 1991. *An outline of Bukiyip grammar* (Pacific Linguistics: Series C). Vol. 113. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Cristofaro, Sonia. 2023. Explaining alienability splits in the use of overt and zero possessive marking: a source-oriented approach. *Linguistics* 61(6). 1613–1641. https://doi.org/doi:10.1515/ling-2022-0034.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: University of Chicago Press.

Croft, William. 2000. Parts of speech as language universals and as language-particular categories. In Vogel, Petra M. & Comrie, Bernard (eds.), *Empirical approaches to language typology*, 65–102. Berlin: Mouton de Gruyter.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, William. 2022. Morphosyntax: Constructions of the World's Languages (Higher Education from Cambridge University Press). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316145289.

Crowell, Thomas H. 1979. *A Grammar of Bororo*. Ithaca: Cornell University PhD dissertation.

Däbritz, Chris Lasse. 2022. *A Grammar of Dolgan: A Northern Siberian Turkic Language of the Taimyr Peninsula* (Grammars and Sketches of the World's Languages). Vol. 18. Leiden: Brill.

Davies, John. 1981. *Kobon: Linguistica Descriptiva Series vol. 3*. Amsterdam: North-Holland.

Derbyshire, Desmond C. 1979. *Hixkaryana syntax* (Lingua Descriptive Series). Amsterdam: North Holland.

Di Garbo, Francesca & Napoleão de Souza, Ricardo. 2023. A sampling technique for worldwide comparisons of language contact scenarios. *Linguistic Typology* 27(3). 553–589. https://doi.org/doi:10.1515/lingty-2022-0005.

Dickens, Patrick. 1992. *Juěrticalline'hoan Grammar*. Windhoek: Nyae Nyae Development Foundation.

Dunn, Michael J. 1999. *A Grammar of Chukchi*. Canberra: Australian National University PhD dissertation.

Emkow, Carola. 2006. *A grammar of Araona, an Amazonian language of Northwestern Bolivia*. Melbourne: LaTrobe University PhD dissertation.

Enrico, John. 2003. *Haida Syntax* (Studies in the Anthropology of North American Indians). Lincoln: University of Nebraska Press.

Evans, Nicholas. 1995. *A grammar of Kayardild* (Moutin Grammar Library 15). Berlin: Mouton de Gruyter.

Evans, Nicholas. 2000. Word classes in the world's languages. In Booij, Geert E. & Lehmann, Christian & Mugdan, Joachim (eds.), *Morphology: a Handbook on Inflection and Word Formation*, vol. 1, 708–732. Berlin: Mouton de Gruyter.

Facundes, Sidney da Silva. 2000. *The Language of the Apurinã People of Brazil (Maipure/Arawak)*. Buffalo: University of New York at Buffalo PhD dissertation. http://hdl.handle.net/11858/00-001M-0000-0012-9A4A-9.

Fenwick, Rohan S. H. 2011. *A Grammar of Ubykh* (LINCOM Studies in Caucasian Linguistics). München: LINCOM.

Fleck, David W. 2003. *A Grammar of Matses*. Houston: Rice University PhD dissertation. http://scholarship.rice.edu/handle/1911/18526.

Forker, Diana. 2020. *A grammar of Sanzhi Dargwa*. Berlin: Language Science Press.

Fortescue, Michael. 1984. *West Greenlandic* (Croom Helm Descriptive Grammars). London: Croom Helm.

Frajzyngier, Zygmunt & Krech, Holly & Mirzayan, Armik. 2002. Motivation for copulas in equational clauses. *Linguistic Typology* 6(2). 155–198.

Gary, Judith Olmsted & Gamal-Eldin, Saad M. 1982. *Cairene Egyptian Colloquial Arabic*. London, Amsterdam: Croom Helm.

Gaved, Timothy. 2020. *A grammar of Mankanya*. Leiden: Leiden University PhD dissertation.

Goddard, Cliff. 1985. *A Grammar of Yankunytjatjara*. Alice Springs: Institute for Aboriginal Development.

Grinevald, Colette G. 1990. *A grammar of Rama*. Unpublished Manuscript.

Hackstein, Olav. 2003. Apposition and word-order typology in Indo-European. In Bauer, Brigitte L. M. & Pinault, Georges-Jean (eds.), *A Festschrift for Werner Winter on the Occasion of his 80th Birthday*, 131–152. Berlin, New York: De Gruyter Mouton. https://doi.org/doi:10.1515/9783110897722.131.

Haiman, John. 1980. *HUA: A Papuan Language of the Eastern Highlands of New Guinea* (Studies in Language: Companion Series). Vol. 5. Amsterdam: John Benjamins.

Haiman, John. 1983. Iconic and Economic Motivation. *Language* 59(4). 781–819. https://doi.org/10.2307/413373.

Hankamer, Jorge. 1973. Unacceptable Ambiguity. *Linguistic Inquiry* 4(1). 17–68. http://www.jstor.org/stable/4177750.

Harms, Phillip Lee. 1994. *Epena Pedee Syntax* (Summer Institute of Linguistics: Publications in Linguistics). Vol. 118. Arlington: The Summer Institute of Linguistics and the University of Texas at Arlington.

Haspelmath, Martin. 2004. Coordinating constructions: An overview. In Haspelmath, Martin (ed.), *Coordinating Constructions*, 3–39. Amsterdam, Philadelphia: John Benjamins.

Haspelmath, Martin. 2007a. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1). 119–132. https://doi.org/doi:10.1515/LINGTY.2007.011.

Haspelmath, Martin. 2007b. Coordination. In Shopen, Timothy (ed.), *Language typology and syntactic description*, vol. 2, 1–51. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Haspelmath, Martin. 2017. Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft* 36(2). 193–231. https://doi.org/doi:10.1515/zfs-2017-0009.

Haspelmath, Martin. 2023a. Word class universals and language-particular analysis. In van Lier, Eva (ed.), *Oxford handbook of word classes*, 15–40. Oxford: Oxford University Press.

Haspelmath, Martin. 2023b. Types of clitics in the world's languages. *Linguistic Typology at the Crossroads* 3(2). 1–59.

Haspelmath, Martin. 2023c. Defining the word. *WORD*. Routledge 69(3). 283–297. https://doi.org/10.1080/00437956.2023.2237272.

Haspelmath, Martin. 2024a. Construction-functions versus construction-strategies. In Däbritz, Chris Lasse & Budzisch, Josefina & Basile, Rodolfo (eds.), *Locative and existential predication: Forms, functions and neighboring domains*. Berlin: Language Science Press. https://ling.auf.net/lingbuzz/007897 (to appear).

Haspelmath, Martin. 2024b. Nonverbal clause constructions. *Language and Linguistics Compass* (to appear).

Hawkins, John A. 2014. *Cross-Linguistic Variation and Efficiency*. Oxford, New York: Oxford University Press.

Heath, Jeffrey. 1999. *A Grammar of Koyra Chiini: The Songhay of Timbuktu* (Mouton Grammar Library). Vol. 19. Berlin: Mouton de Gruyter.

Heine, Bernd & König, Christa. 2010. *The Labwor language of Northeastern Uganda: a grammatical sketch*. Tokyo: Research Institute for Languages and Cultures of Asia and Africa.

Iwasaki, Shoichi & Ingkaphirom, Preeya. 2005. *A reference grammar of Thai*. Cambridge: Cambridge University Press.

Jagersma, Abraham Hendrik. 2010. *A descriptive grammar of Sumerian*. Leiden: Leiden University PhD dissertation.

Jeanne, Leanne M. 1978. *Aspects of Hopi Grammar*. Cambridge: Massachusetts Institute of Technology PhD dissertation. https://dspace.mit.edu/handle/1721.1/16325.

Kazama, Shinjiro. 2011. Typological notes on copula constructions. *Linguistic Typology of the North* (2). 39–54.

Kluge, Angela. 2017. *A grammar of Papuan Malay*. Berlin: Language Science Press. https://langsci-press.org/catalog/view/78/36/614-1.

Koptjevskaja-Tamm, Maria. 2002. Adnominal possession in the European languages: form and function. *STUF - Language Typology and Universals* 55(2). 141–172. https://doi.org/10.1524/stuf.2002.55.2.141.

Koptjevskaja-Tamm, Maria. 2003. Possessive noun phrases in the Languages of Europe. In Plank, Flans (ed.), *Noun Phrase Structure in the Languages of Europe*, 621–722. Berlin: Mouton de Gruyter.

Krishnamurti, Bhadriraju & Gwynn, John Peter Lucius. 1985. *A grammar of modern Telugu*. Oxford: Oxford University Press, USA.

Langacker, Ronald W. 1977. *Studies in Uto-Aztecan grammar* (An Overview of Uto-Aztecan Grammar). Vol. 1. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Langlois, Annie. 2004. *Alive and kicking: Areyonga teenage Pitjantjatjara* (Pacific Linguistics). Vol. 561. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lawyer, Lewis C. 2015. *A Description of the Patwin Language*. Berkeley: University of California at Berkeley PhD dissertation.

Lee, Jennifer R. 1987. *Tiwi today: A study of language change in a contact situation*. Canberra: Research Institute for Languages and Cultures of Asia and Africa, Australian National University.

Levinson, Stephen C. 2022. *The Papuan Language of Rossel Island*. Berlin, Boston: De Gruyter Mouton. https://doi.org/doi:10.1515/9783110733853.

Levshina, Natalia. 2022. *Communicative Efficiency: Language Structure and Use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108887809.

Linn, Mary Sarah. 2001. *A grammar of Euchee (Yuchi)*. Kansas: University of Kansas PhD dissertation.

Logvinova, Natalia. 2024. Towards a typology of specificational constructions. *STUF - Language Typology and Universals* 77(2). 189–233.

https://doi.org/doi:10.1515/stuf-2024-2007.

Lupardus, Karen Jacque. 1982. *The language of the Alabama Indians*. Kansas: University of Kansas PhD dissertation.

MacKay, Carolyn Joyce. 1999. *A Grammar of Misantla Totanac*. Salt Lake City: University of Utah Press.

Maslova, Elena. 2003. *A Grammar of Kolyma Yukaghir* (Mouton Grammar Library). Vol. 27. Berlin: Mouton de Gruyter.

Mayer, Clemens J. 2021. *A grammar sketch of Sentani*. Leiden: Leiden University master thesis.

McGregor, William. 1990. *A Functional Grammar of Gooniyandi* (Studies in Language: Companion Series). Vol. 22. Amsterdam, Philadelphia: John Benjamins.

McKay, Graham. 2000. Ndjébbana. In Dixon, Robert Malcolm Ward & Blake, Barry (eds.), *Handbook of Australian Languages*, vol. 5, 155–356. Oxford: Oxford University Press.

McPherson, Laura. 2013. *A Grammar of Tommo So* (Mouton Grammar Library). Vol. 62. Berlin, Boston, Berlin: De Gruyter Mouton.

Merlan, Francesca. 1983. *Ngalakan Grammar, Texts and Vocabulary* (Pacific Linguistics: Series B). Vol. 89. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Miestamo, Matti & Bakker, Dik & Arppe, Antti. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296. https://doi.org/10.1515/lingty-2016-0006.

Mithun, Marianne. 1988. The grammaticization of coordination. In Haiman, John & Thompson, Sandra A. (eds.), *Clause combining in grammar and discourse*, 331–359. Amsterdam, Philadelphia: John Benjamins Publishing.

Moroz, George. 2017. Lingtypology: easy mapping for Linguistic Typology. https://CRAN.R-project.org/package=lingtypology

Nedjalkov, Vladimir P. & Otaina, Galina A. 2013. *A Syntax of the Nivkh Language: The Amur dialect* (Studies in Language Companion Series). Vol. 139. Amsterdam, Philadelphia: John Benjamins.

Nichols, Johanna. 1988. On alienable and inalienable possession. In Shipley, William (ed.), *In honor of Mary Haas*, 557–610. Berlin: De Gruyter Mouton.

Nikolaeva, Irina & Tolskaya, Maria. 2001. *A Grammar of Udihe* (Mouton Grammar Library). Vol. 22. Berlin: Mouton de Gruyter.

Novak, Irina & Penttonen, Martti & Ruuskanen, Aleksi & Siilin, Lea. 2022. *Karelian in Grammars: A study of phonetic and morphological variation*. Petrozavodsk: KarRC RAS.

Omda Ibrahim Elnur, Elsadig. 2016. *Major Word Categories in Nara*. Universität zu Köln master thesis.

Osborne, Charles. 1974. *The Tiwi language: grammar, myths and dictionary of the Tiwi language spoken on Melville and Bathurst Islands, northern Australia* (Australian Aboriginal Studies). Vol. 55. Canberra: Australian Inst. of Aboriginal Studies.

Payne, Doris Lander. 1985. *Aspects of the grammar of Yagua: A typological approach*. Los Angeles: University of California at Los Angeles PhD dissertation.

Payne, Thomas E. 1997. *Describing morphosyntax: A guide for field linguists*. Cambridge: Cambridge University Press.

Pharris, Nicholas J. 2006. *Winuunsi Tm Talapaas: a grammar of the Molalla language*. Michigan: University of Michigan PhD dissertation.

Piantadosi, Steven T. & Tily, Harry & Gibson, Edward. 2012. The communicative function of ambiguity in language. *Cognition* 122(3). 280–291. https://doi.org/10.1016/j.cognition.2011.10.004.

Plungian, Vladimir Aleksandrovič. 2011. *Vvedenie v grammatičeskuju semantiku: grammatičeskie značenija i grammatičeskie sistemy jazykov mira [Introduction to grammatical semantics: Grammatical meanings and grammatical systems of the world's languages]*. Moscow: Federal'noe gosudarstvennoe bjudžetnoe obrazovatel'noe učreždenie.

Rijkhoff, Jan & Bakker, Dik. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314. https://doi.org/10.1515/lity.1998.2.3.263.

Rijkhoff, Jan & Bakker, Dik & Hengeveld, Kees & Kahrel, Peter. 1993. A Method of Language Sampling. *Studies in Language* 17(1). 169–203. https://doi.org/10.1075/sl.17.1.07rij.

Romero-Figueroa, Andrés. 1997. *A reference grammar of Warao* (Lincom Studies in Native American Linguistics). Vol. 6. Munich: Lincom Europa.

Sadler, Louisa & Nordlinger, Rachel. 2010. Nominal juxtaposition in Australian languages: An LFG analysis. *Journal of Linguistics* 46. 415–452. https://doi.org/10.1017/S002222670999020X.

Schapper, Antoinette. 2022. *A Grammar of Bunaq*. Berlin, Boston: De Gruyter Mouton. https://doi.org/doi:10.1515/9783110761146.

Schultze-Berndt, Eva. 2000. *Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language*. Nijmegen: Wageningen: Ponsen and Looijen PhD dissertation.

Schultze-Berndt, Eva & Simard, Candide. 2012. Constraints on noun phrase

discontinuity in an Australian language: The role of prosody and information structure. *Linguistics* 50(5). 1015–1058.

Seiler, Walter. 1985. *Imonda, a Papuan language*. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Shaul, David Leedom. 2020. *Salinan Language Studies*. Muenchen: LINCOM GmbH.

Smith, Ian & Johnson, Steve. 2000. Kugu Nganhcara. In Dixon, Robert Malcom Ward & Blake, Barry (eds.), *Handbook of Australian Languages*, vol. 5, 357–507. Oxford: Oxford University Press.

Smith, Kenneth. 1979. *Sedang Grammar: Phonological and Syntactic Structure. Pacific Linguistics, Series C-50*. Canberra: Australian National University.

Sneddon, James Neil & Adelaar, Alexander & Djenar, Dwi Noverini & Ewing, Michael C. 2010. *Indonesian reference grammar*. London: Allen & Unwin.

Snyman, Jannie W. 1970. *An Introduction to the !Xũ (!Kung) Language* (University of Cape Town School of African Studies, Communication No. 34). Vol. 34. Cape Town: Department of African Languages, School of African Studies, University of Cape Town.

Sposato, Adam. 2021. *A grammar of Xong* (Mouton Grammar Library). Vol. 84. Berlin: Mouton.

Spronck, Marie-Stephan (Stef). 2015. *Reported speech in Ungarinyin: grammar and social cognition in a language of the Kimberley region, Western Australia*. Canberra: Australian National University PhD dissertation.

Stassen, Leon. 1997. *Intransitive predication*. Oxford: Oxford University Press.

Stassen, Leon. 2000. AND-languages and WITH-languages. *Linguistic Typology* 1(4). 1–54.

Stevenson, Roland C. 1969. *Bagirmi Grammar* (Linguistics Monograph Series). Vol. 3. Khartoum: Sudan Research Unit, University of Khartoum.

Strom, Clay. 1992. *Retuarã Syntax* (Summer Institute of Linguistics: Publications in Linguistics). Vol. 112. Dallas: The Summer Institute of Linguistics and the University of Texas at Arlington.

Terrill, Angela. 2003. A grammar of Lavukaleve. In *A Grammar of Lavukaleve*. Berlin: De Gruyter Mouton.

Terrill, Angela. 2004. Coordination in Lavukaleve. In Haspelmath, Martin (ed.), *Coordinating Constructions*, 427–443. Amsterdam, Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.58.22ter.

Thompson, Sandra A. & Park, Joseph Sung-Yul & Li, Charles N. 2006. *A Reference*

Grammar of Wappo (University of California Publications in Linguistics). Vol. 138. Berkeley, Los Angeles: University of California Press.

Timberlake, Alan. 2004. *A reference grammar of Russian*. Cambridge: Cambridge University Press.

Todd, Evelyn Mary. 1970. *A Grammar of the Ojibwa Language: the Severn Dialect*. North Carolina: University of North Carolina at Chapel Hill PhD dissertation.

Visser, Eline. 2022. *A grammar of Kalamang* (Comprehensive Grammar Library). Berlin: Language Science Press.

Visser, Leontine & Voorhoeve, Clemens. 1987. *Sahu-Indonesian-English Dictionary and Sahu Grammar Sketch* (Verhandelingen van Het Koninklijk Instituut Voor Tall-, Land- En Volkenkunde). Vol. 126. Dordrecht, Holland: Dordrecht: Foris Publications.

Vydrina, Alexandra. 2017. *A corpus-based description of Kakabe, a Western Mande language: prosody in grammar*. Paris: INALCO PhD dissertation.

Wälchli, Bernhard. 2005. *Co-Compounds and Natural Coordination*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276219.001.0001.

Walker, Neil Alexander. 2020. *A Grammar of Southern Pomo*. Lincoln: Nebraska University Press.

Walters, Josiah Keith. 2016. *A grammar of Dazaga* (Grammars and Sketches of the World's Languages). Leiden: Brill.

Wasow, Thomas. 2015. Ambiguity Avoidance is Overrated. In Winkler, Susanne (ed.), *Language and Communication*, 29–48. Berlin, München, Boston: De Gruyter. https://doi.org/doi:10.1515/9783110403589-003.

Weber, David John. 1989. *A Grammar Huallaga (Huánuco) Quechua* (University of California Publications in Linguistics). Vol. 112. Berkeley, Los Angeles: University of California Press.

Appendix

Appendix A provides genealogical information on the sample languages based on Glottolog 5.0, along with the references consulted. Appendix B includes information on the functions that can be expressed by noun juxtaposition in each language and their corresponding references. In Appendix B, "Yes" indicates that noun juxtaposition can be used for the function, while "–" signifies that the use of noun juxtaposition for that function cannot be found in the indicated sources.

Appendix A: A list of the sample languages.

Area	Language	Family	References
Africa	Tommo So	Dogon	McPherson (2013)
	Ju|'hoan	Kxa	Snyman (1970); Dickens (1992)
	Kakabe	Mande	Vydrina (2017)
	Koyra Chiini	Songhay	Heath (1999)
	Dazaga	Saharan	Walters (2016)
	Egyptian Arabic	Afro-Asiatic	Gary & Gamal-Eldin (1982)
	Nara	Isolate	Omda Ibrahim Elnur (2016)
	Bagirmi	Central Sudanic	Stevenson (1969)
	Labwor	Nilotic	Heine & König (2010)
	Kunama	Isolate	Bender (1996)
	Tima	Kalta-Tima	Alamin Mubarak (2009)
	Mankanya	Atlantic-Congo	Gaved (2020)
Australia	Kugu Nganhcara	Pama-Nyugan (Paman)	Smith & Johnson (2000)
	Nhanda	Pama-Nyugan (South-West Pama-Nyugan)	Blevins (2001)
	Gooniyandi	Bunaban	McGregor (1990)
	Jaminjung	Mirndi	Schultze-Berndt (2000); Schultze-Berndt & Simard (2012)
	Ndjébbana	Maningrida	McKay (2000)
	Pitjantjatjara	Pama-Nyugan (Desert-Nyungic)	Langlois (2004)
	Jiwadja	Iwaidja Proper	Capell (1962)
	Tiwi	Isolate	Lee (1987); Osborne (1974)
	Ngarinyin	Worrorran	Coate & Oates (1970); Spronck (2015)
	MalakMalak	Northern Daly	Birk (1976)

Area	Language	Family	References
	Kayardild	Tangkic	Evans (1995)
	Ngalakgan	Gunwinyguan	Merlan (1983)
Eurasia	Chukchi	Chukotko-Kamchatkan	Dunn (1999)
	Kolyma Yukaghir	Yukaghir	Maslova (2003)
	Sanzhi Dargwa	Nakh-Daghestanian	Forker (2020)
	Xong	Hmong-Mien	Sposato (2021)
	Ubykh	Abkhaz-Adyge	Fenwick (2011)
	Amur Nivkh	Nivkh	Nedjalkov & Otaina (2013)
	Thai	Thai-Kadai	Iwasaki & Ingkaphirom (2005)
	Telugu	Dravidian	Krishnamurti & Gwynn (1985)
	Udihe	Tungusic	Nikolaeva & Tolskaya (2001)
	Dolgan	Turkic	Däbritz (2022)
	Duhumbi	Sino-Tibetan	Bodt (2020)
	Sedang	Austroasiatic	Smith (1979)
North America	Haida	Isolate	Enrico (2003)
	Salinan	Isolate	Shaul (2020)
	Yucatec Maya	Mayan	Bolles & Bolles (2014)
	Wappo	Uki-Wappo	Thompson et al. (2006)
	Yuchi	Isolate	Linn (2001)
	Alabama	Muskogean	Lupardus (1982)
	Misantla Totonac	Totonacan	MacKay (1999)
	Rama	Chibchan	Grinevald (1990)
	Severn Ojibwa	Algic	Todd (1970)
	Hopi	Uto-Aztecan	Langacker (1977); Jeanne (1978)
	Southern Pomo	Pomoan	Walker (2020)
	Patwin	Wintuan	Lawyer (2015)
Papunesia	Ulwa	Keram	Barlow (2023)
	Yélî Dnye	Isolate	Levinson (2022)
	Bunaq	Timor-Alor-Pantar	Schapper (2022)
	Hua	Nuclear Trans New Guinea (Kainantu-Goroka)	Haiman (1980)
	Indonesian	Austronesian	Sneddon et al. (2010)
	Sentani	Sentanic	Mayer (2021)
	Kobon	Nuclear Trans New Guinea (Madang)	Davies (1981)
	Lavukaleve	Isolate	Terrill (2003); Terrill (2004)

Area	Language	Family	References
	Imonda	Border	Seiler (1985)
	Bukiyip	Nuclear Torricelli	Conrad & Wogiga (1991)
	Sahu	North Halmahera	Visser & Voorhoeve (1987)
	Kalamang	West Bomberai	Visser (2022)
South America	Hixkaryana	Cariban	Derbyshire (1979)
	Matses	Pano-Tacanan (Panoan)	Fleck (2003)
	Apurinã	Arawakan	Facundes (2000)
	Warao	Isolate	Romero-Figueroa (1997)
	Araona	Pano-Tacanan (Tacanan)	Emkow (2006)
	Huallaga Quechua	Quechuan	Weber (1989)
	Epena	Chocoan	Harms (1994)
	Yagua	Pebe-Yagua	Payne (1985)
	Sanuma	Yanomamic	Borgman (1990)
	Paumarí	Arawan	Chapman & Derbyshire (1991)
	Bororo	Bororoan	Crowell (1979)
	Retuarã	Tucanoan	Strom (1992)

Appendix B: All data of the sample.

Language	Predication	Possession	Conjunction	References
Tommo So	–	Yes	–	McPherson (2013: 339-349; 183; 211-213)
Ju|'hoan	–	Yes	–	Snyman (1970: 127); Dickens (1992: 17; 33)
Kakabe	Yes	Yes	–	Vydrina (2017: 74; 92; 118)
Koyra Chiini	–	Yes	–	Heath (1999: 143-148; 84-85; 113-116)
Dazaga	Yes	–	–	Walters (2016: 143-147; 63; 173-177)
Egyptian Arabic	Yes	Yes	–	Gary & Gmal-Eldin (1982: 61; 48-49; 36-37)
Nara	Yes	Yes?	–	Omda Ibrahim Elnur (2016: 73; 39; 49)
Bagirmi	Yes	Yes	–	Stevenson (1969: 163; 57; 182)
Labwor	Yes	Yes	–	Heine & König (2010: 29-30; 61; 98)
Kunama	Yes	–	–	Bender (1996: 41-43; 18-19; 23)
Tima	Yes	Yes	–?	Alamin Mubarak (2009: 202; 130-131; 96)
Mankanya	Yes (restricted)	Yes	–	Gaved (2020: 124-125; 136; 104)
Kugu Nganhcara	Yes	–	Yes	Smith & Johnson (2000: 389, 418; 428; 434)
Nhanda	Yes	–	Yes	Blevins (2001: 46, 62, 66, and ff.; 66, 57; 133-134)
Gooniyandi	Yes	Yes	Yes	McGregor (1990: 294-302; 252-253, 261; 284-285)
Jaminjung	Yes	–	Yes	Schultze-Berndt (2000: 109; 63-69, 184-185, and ff.); Schultze-Berndt & Simard (2012: 1052)
Ndjébbana	Yes	Yes	Yes	McKay (2000: 292; 195; 306-307)
Pitjantjatjara	Yes	Yes	Yes	Langlois (2004: 85; 84); Bowe (1990: 43)
Jiwadja	Yes	Yes	Yes	Capell (1962: 164; 155; 160)
Tiwi	Yes	Yes	Yes	Lee (1987: 285-286); Osborne (1974: 74); Lee (1987: 230-231)
Ngarinyin	Yes	Yes	–	Coate & Oates (1970: 66); Spronck (2015: 39; 38)
MalakMalak	Yes	Yes	Yes	Birk (1976: 126, 153; 106; 122, 148)

Language	Predication	Possession	Conjunction	References
Kayardild	Yes	Yes	Yes	Evans (1995: 313-314; 247-249; 250)
Ngalakgan	Yes	Yes	Yes	Merlan (1983: 57-61; 82; 148)
Chukchi	Yes	–	–	Dunn (1999: 83, 317-318; 149-151; 172-174)
Kolyma Yukaghir	–	Yes	Yes	Maslova (2003: 437-441; 290; 316-318)
Sanzhi Dargwa	Yes (restricted)	–	Yes	Forker (2020: 429-430; 574-575; 506)
Xong	Yes (restricted)	Yes	Yes	Sposato (2021: 402; 389; 395)
Udykh	Yes	–	–	Fenwick (2011: 155-156; 46-51; 62)
Amur Nivkh	Yes	Yes	–	Nedjalkov & Otaina (2013: 37-38; 1, 9, 14; 56-58)
Thai	Yes	Yes	–	Iwasaki & Ingkaphirom (2005: 228-229; 65-66; 10, 171-172)
Telugu	Yes	–	–	Krishnamurti & Gwynn (1985: 308-310; 76, 82; 325-327)
Udihe	Yes	Yes (restricted)	Yes	Nikolaeva & Tolskaya (2001: 608-609; 785-786; 647-648)
Dolgan	Yes	–	Yes	Däbritz (2022: 362; 157-169; 320)
Duhumbi	Yes	–	–	Bodt (2020: 395-397; 281; 594-595)
Sedang	Yes	Yes	–	Smith (1979: 116-117; 76-77; 154)
Haida	–	Yes	Yes	Enrico (2003: 211-212, but 135-136; 706, 709; 1079)
Salinan	Yes	–	–	Shaul (2020: 83; 80; 106)
Yucatec Maya	Yes	–	–	Bolles & Bolles (2014: 21; 20; 65)
Wappo	Yes (restricted)	Yes	Yes	Thompson et al. (2006: 103; 15-16; 22-23)
Yuchi	Yes	–	Yes	Linn (2001: 416-417; 383-390, 398; 511)
Alabama	Yes	–	–	Lupardus (1982: 217; 94-100; 239-240)
Misantla Totonac	Yes	–	–	MacKay (1999: 404-405; 347-352; 436)
Rama	Yes	Yes	Yes?	Grinevald (1990: 96, 130; 94; 239)

Language	Predication	Possession	Conjunction	References
Severn Ojibwa	Yes	–	Yes	Todd (1970: 79; 32-34; 41)
Hopi	Yes	–	Yes	Langacker (1977: 40); Jeanne (1978: 112-125); Langacker (1977: 160)
Southern Pomo	–	Yes (restricted)	Yes	Walker (2020: 170-171, 243, 270; 154; 335)
Patwin	–	–	Yes	Lawyer (2015: 294-295; 92, 142-148; 190)
Ulwa	Yes	–	Yes	Barlow (2023: 320; 175-179; 353)
Yélî Dnye	Yes	–	Yes	Levinson (2022: 284-286; 165; 163)
Bunaq	Yes	–	Yes	Schapper (2022: 131-132; 329; 225)
Hua	Yes	Yes	–	Haiman (1980: 345; 366; 249-)
Indonesian	Yes	Yes	–	Sheddon et al. (2010: 242; 148-150; 347)
Sentani	Yes	Yes	–	Mayer (2021: 63-64; 45; 39)
Kobon	Yes	Yes	Yes	Davies (1981: 41-42; 57; 72)
Lavukaleve	Yes	–	Yes	Terrill (2003: 240; 93-97); Terrill (2004: 431)
Imonda	Yes	–	–	Seiler (1985: 154; 62-63; 68-69)
Bukiyip	Yes	–	Yes	Conrad & Wogiga (1991: 90-91; 65; 63-64)
Sahu	Yes	–	Yes (restricted)	Visser & Voorhoeve (1987: 59; 53-54; 54)
Kalamang	Yes	–	Yes	Visser (2022: 293; 217-227; 146, 185)
Hixkaryana	Yes	–	Yes	Derbyshire (1979: 36-37; 69-70; 45-46)
Matses	–	Yes	Yes	Fleck (2003: 944-950; 764; 805)
Apurinã	Yes	Yes	Yes	Facundes (2000: 504; 152-153; 426)
Warao	Yes	–	Yes	Romero-Figeroa (1997: 11, 38; 44-45, 90-91; 12-13)
Araona	Yes	Yes	Yes?	Emkow (2006: 407-408; 41-42; 690)
Huallaga Quechua	Yes	–	Yes	Weber (1989: 24; 54-55; 20, 347-348)
Epena	Yes	Yes	–	Harms (1994: 33-34; 49-52; 55)

Language	Predication	Possession	Conjunction	References
Yagua	Yes	Yes	Yes	Payne (1985: 57-58; 155-156, 83-86; 97, 83-86)
Sanuma	Yes	Yes (restricted)	–	Borgman (1990: 20-21; 127; 34-35)
Paumarí	Yes	–	–	Chapman & Derbyshire (1991: 168-169; 256-259; 189)
Bororo	Yes	–	–?	Crowell (1979: 38-39; 214-217; 241-245)
Retuarã	Yes	Yes	–	Strom (1992: 129; 5, 48; 39)

CONTACT

efforts.0213@gmail.com

Linguistic Typology at the Crossroads 4-2 (2024): 273-289 ISSN 2785-0943

Reflections on the "ad hoc categories"

PAOLO RAMAT

ACCADEMIA NAZIONALE DEI LINCEI - ROMA

Submitted: 24/03/2024 Revised version: 31/07/2024
Accepted: 31/07/2024 Published: 23/01/2025

Abstract

This article is a reflection on the concept of *ad hoc categories* (AHCs) as developed in a copious number of recent publications. The article refers to well-known concepts such as *prototype*, and theoretical frameworks such as *cognitivism*, and *construction grammar*, which are shortly presented in section 1 inasmuch they may concern the discussion of AHCs and are preliminary to such a discussion. Section 1 deals with the definition(s) of category, section 2 presents the notion of AHC, section 3 deals with different types of AHC, and section 4 discusses some problems connected to this notion and its possible limits. Section 5 is the conclusion that can be drawn from the previous reflections.[1]

Keywords: category; categorization; prototype; general extenders; collective nouns; languaging activity.

1. The notion of category

According to the on-line Vocabolario Treccani of the Italian Encyclopaedia a category is a "partizione nella quale si comprendono individui o cose di una medesima natura o di un medesimo genere" ('A division that contains individuals or things having the

[1] Since this paper deals with general problems concerning AHCs, it is not based on a particular corpus. The examples in the text are quoted from the discussed literature. I have kept in the glosses of the examples the original glossing of the Authors. Consequently, there may be some inconsistency in the glossing system. I wish to thank two anonymous reviewers for their insightful, helpful observations.

same nature or same genre'). This is the traditional, rigid definition, according to which an X belongs or does not belong to a given category. A look at the standard monolingual dictionaries of our (Western) tradition confirms Treccani's definition. The French dictionary Larousse has the following definition of *catégorie*: "Ensemble de personnes ou de choses de même nature" ('Ensemble of people or things which have the same nature') and provides a list of synonyms: *espèce - famille - genre - groupe - sorte*. Thereafter, in addition to different sorts of categories such as *Boucherie*, *Philosophie*, *Sports*, *Logique*, *Mathématiques*, a paragraph is also dedicated to Linguistics: "Unité de classement grammatical qui peut correspondre soit à la notion de classe (catégories du nom, de l'adjectif, du déterminant, du verbe, etc.), soit à la notion de constituant (catégorie du syntagme nominal, du syntagme verbal, etc.), soit aux modifications que peuvent subir les classes (catégories du nombre, du genre, du temps, de la voix, du mode, etc.)". In the Digitales Wörterbuch der deutschen Sprache (*DWDS*) we find the same reference to people or things: "Gruppe, in die jemand oder etwas eingeordnet wird" ('Group, where someone or something is inserted').[2] The online Oxford English Dictionary (*OED*) (with reference to Linguistics) states the following: "A class, or division, in any general scheme of classification". Similarly, the Diccionario de la lengua Española (*DLE*, Real Academia Española) gives a general definition: "Cada una de las clases o divisiones establecidas al clasificar algo" ('Every class or division established when classifying something'), further referring to grammatical categories (e.g. gender and number) and *clases de palabras* (e.g., noun and adjective). While Treccani and Larousse do not use the verb "classify" in their definitions, the *OED* and the *DLE* seem a bit tautological: a class is the result of a classification. This is obviously correct, but the question remains: what is a classification? In other words: a category is the product of categorization, but we have to define how we accomplish the categorizing operation.

It should be noted that the above-mentioned dictionaries (as well as other standard dictionaries) dedicate a paragraph to linguistic categories, considering them mostly from the morphological or morphosyntactic viewpoint. The already cited Vocabolario

[2] One out of the many instances of the word Kategorie reported in the *DWDS* is for our discussion particularly relevant: Schneider 1965: s.9: "Möbel, Häuser, Kleider, Küchengeräte usw. gehören in die **Kategorie** der dauerhaften Güter, während Streichhölzer, Zigaretten, Tinte usw. zur **Gruppe** der Verbrauchsgüter zählen" ('Furniture, houses, dresses, tools for the kitchen, etc. belong to the **category** of lasting objects, whereas matches, cigarettes, ink, etc. belong to the consumer goods **group**'). I'll come back to such a distinction in section 5.

Treccani (s.v. *categoria linguistica*) mentions the categories SN = sintagma nominale, SV = sintagma verbale (in English NP and VP, respectively), N = nome, V = verbo, and Art = articolo as symbols used to represent a specific category. To the classical categories (or parts of speech: PoS) N, V, and Art, SV and SN are added, which pertain to syntax. The same holds for Larousse's distinction between *classe* and *constituent*.

1.1. Categories and categorizations

Linguists have always been well aware that their categories are not completely black or white and that there exist elements that are difficult to classify. Consider, for instance, the participle, whose name says that it *partem capit*, *participates* in the verbal and adjectival nature. Moreover, different PoS may share the property A but not the property B: for instance, the categories of participle and gerund share in Romance languages the feature [deverbal] but differ as to the feature [adjectival] vs. [adverbial].

With reference to mathematics the Vocabolario Treccani states: "Affinché un insieme possa ritenersi definito è necessario che ne siano assegnati gli elementi, oppure che per essi sia assegnata una proprietà caratteristica, cioè un criterio per decidere se un certo oggetto è o no elemento di un certo insieme". ('In order to consider an ensemble as defined, it is necessary that its elements be assigned or a characteristic property be determined for them, i.e., a criterion capable of deciding whether a given object is or is not an element of the ensemble'). The "characteristic property" ('proprietà caratteristica') is the deciding point, but it is implicitly admitted that an element belonging to the category A because it has the *proprietà caratteristica* of A may also have other properties. This is particularly true for linguistics (see above the example of participle and gerund).

However, in the last decades of the twentieth century, the introduction of the concept of *prototype* has further weakened the boundaries of the traditional categories, not only in linguistics but everywhere the concept of category can be applied. As stated by Mauri et al. (2021: 30), "categorization appears to be often instrumental to intersubjective aims, such as mutual agreement and the general management of the speakers' reciprocal positioning"; "speaker and hearer are mutually and contemporarily involved in the identification of the category members and the category boundaries, recurring to exemplification along a progressive zooming-in movement".

At the same time, both Cognitive Linguistics that analyses linguistic expressions according to the cognitive processes which generate them, and Construction Grammar, where the starting point of every linguistic analysis must be that all linguistic expressions are a combination of different constructs which together specify the form and the meaning, have largely widened the horizons beyond the traditional categories of the parts of speech (*partes orationis*) such as Noun or Verb.[3] Texts and sentences constitute the main aspects of analysis. Such a widening is strictly connected with the notion of which, in turn, represents a crucial enlargement of the category concept.

Prototype theory admits that along with eagles, sparrows and swallows also penguins, ostriches, and the now extinct dodos also belong to the category BIRD, although they are (were) unable to fly, where FLYING may represent the most important characteristic of birds. BIRD is a taxonomic, "natural category",[4] endowed with core representative and less representative elements. Moreover, it is possible to have "not-natural categories" which assemble material things or abstract concepts according to the co(n)textual situation. For instance, in a hunt scenario the hunters can speak of foxes, pheasants and wild boars as an ad hoc category (let us name it PREY ANIMALS), strictly bound to the particular situation of hunting in a particular location inhabited by foxes, pheasants and wild boars (thus, not in Arabia nor in Greenland). A category is the end-product of a bottom-up exemplar-driven procedure – let us refer to it as categorization – which collects elements sharing some relevant properties. Birds are characterized by the capacity to fly, lay eggs, etc. Once the category BIRD has been defined via the cognitive procedure that recognizes peculiar similarities between (mental) objects, new members can be added via a top-down procedure: and this concerns not only "regular" birds as condors or parrots but also "less regular" ones such as penguins or dodos (see Sammarco 2021: 234).

2. The "ad hoc categories"

The concept of "ad hoc category" (henceforth AHC) was formulated by the psychologist and cognitive scientist Lawrence W. Barsalou in 1983 and published, not

[3] In the frame of Cognitive Grammar Langacker (1987: 377-396 and 409-411) considers a category as a *network* of schemas.

[4] On the notion of "natural category" see Eleanor Rosch's fundamental writings (Rosch 1973; Rosch et al. 1976; etc.).

by chance, in the journal *Memory and Cognition*. The concept was immediately used in linguistics. Barsalou developed his ideas in many books and papers, up to his chapter on "Categories at the interface of cognition and action" which represents, so to speak, a summa of his writings on the subject (Barsalou 2021). After Barsalou's milestone writings a flood of articles appeared in the '90s and in the first two decades of the present century. To quote just the most significant publications, *Folia linguistica historica* issued a special volume edited by Caterina Mauri & Andrea Sansò (vol. 39, 2018), titled "Linguistic strategies for the construction of ad hoc categories: synchronic and diachronic perspectives". The journal *Language Sciences* published a volume (No 81, 2020) edited by Caterina Mauri & Andrea Sansò with the title "Ad hoc categorization and language: the construction of categories in discourse". A book edited by Caterina Mauri, Ilaria Fiorentini & Eugenio Goria and published by John Benjamins appeared 2021: "Building Categories in Interaction: Linguistic resources at work". Other papers are scattered throughout linguistic journals and books, often written by the same authors who contributed to the above-mentioned publications.

As is often the case, new ideas are adopted with enthusiasm and sometimes extended beyond their original limits. In what follows I attempt to take up a stance on the issues which have been discussed thus far in the literature to date.

3. Different types of the "ad hoc categories"

Barsalou's standard and most comprehensive definition of an AHC is as follows (2010: 86): "An ad hoc category is a novel category constructed spontaneously in achieve a goal relevant in the current situation". AHCs are, for instance, "ways to get from San Francisco to New York", "foods not to eat on a diet" etc., which appear to be constructed spontaneously when the co(n)textual situation suggests/needs them. Consequently, an AHC, as Mauri (2017: 299; 2021: 29) states, is the output of a bottom-up, goal-driven and context-dependent process abstracting from specific exemplars (e.g. foxes, pheasants, and wild boars) in a particular situation.

The difference between category and categorization is crucial. Mihatsch (2018: 148) correctly writes that "[t]he term 'categorization' refers to the assignment of a category to an individual": this is correct, although I would prefer assigning an individual to a category. However, "categorization" may also mean the creation of a category via the bottom-up procedure previously alluded to. Given that AHCs are highly context- and situation-dependent and people construct them to achieve their

ad hoc communicative goals, the question arises: are there limits to AHCs or, as Barsalou (1999: 578) maintains, the number of human categories is essentially infinite? To give a plausible answer to this question, it is necessary to have recourse to the linguistic forms AHCs may take on. Mauri (2017: 300) states that there are "non-random correlations between specific morphosyntactic properties and specific ways of abstracting the categories". For instance, in the text

(1) We are in Rome for the weekend. We have plenty of things to do, you know: [visit the Colosseum, stroll through the Gardens of the Villa Borghese, go to the Trevi fountain, **and so on**…] everything in two days!

the hearer understands that the monuments mentioned are only a part of an AHC which could be dubbed as MONUMENTS TO VISIT IN ROME. The cue for such interpretation is the general extender (see fn. 6) *and so on* and the list functions as an exemplification of the ad hoc invented category, which is created for a particular situation. Let us consider another example proposed by Barsalou (2021) as THINGS TO PACK IN A SUITCASE (properly a "goal-derived category"). If one limits oneself to mention *toothpaste, toothbrush, socks* and *pants* without finishing the list by *etc.* or *and so on,* or at least by a suspensive tone, the interlocutor is entitled to ask: "Any other thing?".

In other words, a "more-to-come" element[5] (*and the like, and so on, things like that,* etc.) indicates the creation of an AHC. The list that forms, so to say, the *incipit* of the AHC (in ex. (1) the Colosseum, the Gardens of the Villa Borghese, the Trevi fountain) is characterized by the "syntagmatic concatenation of two or more units of the same type" (Masini et al. 2018: 50), whereby "same type" is to be intended as the "syntactic and functional same type", since we have seen that an AHC can contain elements of very different nature, like foxes and pheasants. In their introduction to Mauri et al. (2021: 2), Mauri, Fiorentini, & Goria give a list of "special strategies" used to build AHCs: marked prosodic and morphological patterns, reduplication, associative and similative plurals, list constructions, exemplification, and general extenders.[6] In other words, linguistic data can reveal the process of category construction: linguistic

[5] Cp. Goria & Masini 2021: 75.

[6] The "general extenders" such as *and so on, etc., something like that* represent a strategy of abstraction done by the speaker that may include also non-specific items: see Mauri & Giacalone Ramat 2015 (particularly on Japanese *–tari*), Mauri & Sansò 2018. Moravcsik 2020 offers a taxonomy of AHCs expressed by plurals (for instance the 'similative plurals' as Telugu (tel; Dravidian, South Dravidian) *puli-gili* 'tigers and such' (*puli* 'tiger'). On "echo words" see below, fn. 8).

means/strategies are needed to form a category– and more specifically an AHC, which indicates an ensemble of similar/analogous material or conceptual things– be it a connective item like the Japanese *toka* in

(2) Japanese (Barotto 2018: 44)
Kōhī toka kōcha toka iroirona mono-ga arimashita
coffee TOKA tea TOKA various$_{ADJ}$ hing-NOM exist$_{POL:PAST}$
'There were various things such as coffee and tea.'

or the "echo compounds" that are formed, as in Lezgian,[7] "by reduplicating nouns in such a way that the onset of the first syllable of the second member is replaced by *m-*. The meaning of such *N m-N* compounds is "*N and similar things"* (Haspelmath 1993: 109; my emphasis), so that we get *sik'~mik'* 'fox and other wild animals' (*sik'* 'fox').

Both the "echo compounds"[8] and the *toka*-connectives are categorizing tools, i.e. "categorization triggers" (see Mauri & Sansò 2018b: 1). Goria & Masini (2021: 78) distinguish between "categorizing" (or "category-building") lists and "lists that implicitly rely on some presupposed category" as in Rosch's "natural categories" like ANIMALS or STARS (see fn. 4).

As we have seen (cp. ex. (2)), not all AHC markers must systematically occur at the end of the list completer slot like *and so on* or *and the like*. Italian *tipo, che so* (properly, a one-word: [ke's:o], just as English *dunno*; see fn. 10) 'I don't know', and French *genre* introduce the AHC. The following example is drawn from a corpus of spoken Italian mostly used by internet-newsgroups as reported by Lo Baido (2018: 80):

(3) *Mi ha chiesto cose **tipo** Moby, Eminem, Saggy **insomma** che non siano solo dance un po' misto ecco.*
'(S)He asked me for things **tipo** Moby, Eminem, Saggy **in sum** that are not only dance, a little bit mixed, I mean.'

[7] lez; Nakh-Daghestanian, Lezgic.

[8] The echo-word construction, a non-canonical reduplication, is attested in various languages: see, for instance, Turkish ***Dergi mergi*** *okumuyor*/Newspapers M:ECHO read:NEG:PRES , '(S)He does not read newspapers and the like' (Stolz 2018: 248; see also Stolz 2003/04: 11). Magni (2018: 204) speaks for such cases of "echo twin strategy". Kallergi (2015: 18 -as well as Haspelmath 1993: 109) considers this construction not only as signalling vagueness but also somehow deprecative, pejorative (and this is quite understandable as the consequence of vagueness, uncertainty).

A sentence like **Mi ha chiesto cose Moby, Eminem, Saggy* ***tipo****, insomma che non siano solo dance, un po' misto ecco,* showing *tipo* at the end of the list, would be impossible. *Tipo* and its equivalents in other languages (French *genre*, *espèce,* Spanish and Portuguese *tipo*, Russian *tipa* (GEN), English *kind, type, sort*)[9] usually introduce the AHC. One element is sufficient to create the AHC:

(4) *Una piccola polemica "elegante"* ***tipo*** *Accademia della Crusca* (L. Romano, 1969, *Le parole tra noi leggere*, quoted by Voghera 2013a: 296).

The mention of the Accademia della Crusca, a well-known and precisely defined object, is sufficient to create the category ACADEMIES WITH *ELEGANT* DISCUSSIONS. The Accademia della Crusca is the exemplifying placeholder which frames the conceptual space it belongs to (cp. Lo Baido 2018: 86).

3.1. On general extenders

Contrary to *tipo* and its above mentioned equivalents such as *kind*, *genre,* et sim., the general extenders *che so,*[10] *I don't know, que sais-je* et sim., are (or, at least, originally were) sentences *per se* and though sometime used inside the AHC list (see the example of *chi sacciu*, fn. 9), they usually close the AHC, often using a suspensive tone:

(5) a. *C'est comme l'entente sur les soins de santé ou* ***que sais-je***
'It is like the health care deal or whatever.'
b. *Il aurait pu m'envoyer une note, un accord* ***que sais-je?...***
'He could have sent me a note, an agreement, whatever.'

[9] On Italian *tipo* and related forms see Voghera 2012; 2013a; 2013b.

[10] See De Mauro 2000: 444, (s.v. *che*) "e altre cose dello stesso genere: *aveva tutte le qualità era brava, bella, gentile* ***e che so io***", lit. 'she had all the good qualities: she was skilful, beautiful, courteous **and what I know**'. Lo Baido 2023 has studied the corresponding Sicil. *chi sacciu,* lit. 'what do I know', i.e. 'I don't know, I dunno'. She underlines the "basso grado di coinvolgimento assertivo al fine di dichiarare lo status ipotetico ed esemplificativo di alcuni *items*" (p.140: 'the low commitment of the speaker in order to underline the hypothetical and just exemplifying role of some items'); *ci poi regalare* ***chi sacciu*** *na penna, un portachiavi bonu* 'you can give him/her as gift, **chi sacciu** (what I dunno) a pen, a fine key chain', Lo Baido, loc.cit, ex. (31): the pen and the fine key chain are representative of the open list that forms the AHC 'THINGS TO BE GIVEN AS GIFT (IN A PARTICULAR OCCASION)'. Lo Baido (p.c.) adds that *chi sacciu* may occur also at the end of the sentence, just as *che so, I don't know > I dunno, and so on, que sais-je?*. On *dunno* and similar forms see further fn. 12.

Similarly to Accademia della Crusca in (4), in (5a) the health care deal is sufficient to represent a category (say STATE MEASURES FOR THE CITIZENS); in (5b) there are multiple elements but a sentence like *Il aurait pu m'envoyer une note, que sais-je?...* would also be fine.

We may conclude that it is not the number of the list members that creates an AHC. However, it is rather rare finding AHC closing expressions (*and so on*, etc.) preceded by just one element as in (6) below. Moreover, not all lexemes have the same capacity to construct an AHC. In my opinion, the sentence quoted by Mauri & Sansò (2018b: 26) does not constitutes a good example of an AHC:

(6) *It was some sort of chessboard, you know, not a real chessboard, more like a large, decorated disk, a shield, something like that. A round chessboard–like object.*

The speaker refers here to a single object (s)he has problems defining. Contrary to the Accademia della Crusca in (4) and even to *une note, un accord* in (5b), it cannot be ascribed to a specific category nor represent the starting point of a newly *ad hoc* created category ('chessboards', 'shields', 'round objects'?...). On the shortcomings of categorizing on the basis of lexical items see Barotto 2018: 39.

4. Some distinctions among the "ad hoc categories"

The question to be discussed at this point is: are all the previous examples really AHCs? Mauri & Sansò (2018a: 70) make the important distinction between *insiemi* (ensembles) and *classi* (classes) The former are represented, among others devices, by the associative and collective plurals such as the Hungarian suffix *–ék: Jánosék* 'Janos and his relatives', or Japanese *-tachi* in *Tanakatachi* 'Tanaka and people associated with him' (Mihatsch 2018: 151; Moravcsik 2020: ex. (12)). Classes may use disjunctive connectives like *or* as in

(7) *I came to class but they have a bomb threat* ***or something*** (ex. (13) in Mauri & Sansò 2018a).

Clearly, *-ék* and *–tachi* are not goal-derived AHCs in Barsalou's sense (see above, section 3), whereas the general extender *or something* in (7) builds the class EVENTS THAT KEEP STUDENTS OUT OF THE CLASS.

4.1. Ad hoc categories and collective nouns

A further point that can help us to better understand the concept of AHC deserves to be underlined in the frame of the general discussion that appears in recent publications: the collective nouns (or "aggregates") like Italian *fogliame* 'foliage', *vasellame* 'tableware' *ciarpame* 'rubbish, junk', studied by Magni (2018) are not AHCs, even less goal-derived AHCs. They are regular entries of the Italian dictionaries, not bound to a particular situation. The *–ame* suffix can also be attached to proper nouns of celebrities or well-known politicians to denote the set of persons, ways of acting, situations whose pivot is the proper noun: *Berlusconi* → *berlusconame,* is yet a nonce-noun[11] strictly bound to the popularity of Berlusconi. It might well be that it be registered in the future in some (historical) Italian dictionary. At the present moment I would say that *berlusconame* is – or, better, has been– on the way of becoming an AHC.

As for the Italian nouns with the collectivizing suffix *–ume* (*marciume* 'rot, rottenness' (< *marcio* 'rotten'), *sudiciume* 'dirt, filth' (< *sudicio* 'dirty', and the like), we observe that a sentence as

(8) *Si vede dappertutto sudiciume* ***e così via (/e simile)*** *(*or other AHC-markers).
'One can see everywhere dirt **and so on**' (my own example).

would sound very strange, since *sudiciume* does not constitute a category, but just a state of affairs or an ensemble of things that are dirty (but not *dirty and so on*!). Collective nouns (*aggregates*) can be specified: e.g. *sudiciume* may be the cover noun for *gums, stubs, empty cans*, etc. Consequently*, sudiciume* may be for the speaker the starting point for constructing an AHC as in *I saw in that rave party just sudiciume, marciume, sfasciume* ['junk'] ***and things like that***. Mauri & Sansò (2018b: 23; my italics) write: "Collective and aggregate markers are among the *morphological strategies* used to encode ad hoc categorization across languages". It is, however, important to repeat that aggregates, collectives, like the *pluralia tantum* (e.g. Lat.

[11] As already said in fn.8, Haspelmath and Kallergi note the generally pejorative connotation of some *ad hoc* categorization triggers like 'echo-words' (mentioned in section 3). The same holds for the ephemeral creations like *berlusconame* and the *–aglia* collectives as *salvinaglia* 'people and/or affairs around the right-wing politician Matteo Salvini', that is analogically formed on *marmaglia* 'riff-raff', *gentaglia* (<*gente*) 'rabble, scum', *teppaglia* (<*teppa*) 'hooligans' etc., i.e. on pejoratives which are completely lexicalized -along with non-pejoratives as *boscaglia* (<*bosco*) 'boscage 'or *nuvolaglia* (<*nuvola*) 'mass of scattered clouds' (Magni 2018: 212; Arcodia & Mauri 2020).

deliciae ‘delight’*, divitiae* ’richness’) are not AHCs, i.e. categories created under particular circumstances. Intrinsically, *fogliame,* which, as said above, belongs to the Italian lexicon, denotes *per se* an amount of leaves, without any further specification, no matter whether the leaves are from a fruit tree, an oak or a pine. On the other hand, when the speaker alludes to THINGS TO PACK IN A SUITCASE the hearer expects that the speaker specifies which objects have to be put in the suitcase as there is no collective noun referring to such things.

Furthermore, we have to distinguish between grammatical(ized) tools like collectivizing suffixes (as It. *–ame*, *–ume*) and spontaneously, mainly conversationally, created expressions such as *and things like that,* or *I dunno, que sais-je, che so* (*io*)*, was weiß ich, quién sabe, ne znayu*: these expressions are stereotyped and belong to the common language use (‘Sprachschatz’), but they are not grammatical tools. They can be used to signal the creation of AHCs. In short, there is not only a division between “natural categories” and AHCs, but the latter are further divided into morphological and conversational building strategies of languaging.[12]

4.2. The languaging activity

Inglese & Geupel (2018: 228 and 236) present sentences with a list of examples, introducing, following Mauri 2017, the threefold division in sets, classes, and frames with the following examples:

(9) a. *I need flour, milk, yeast* ***and so on*** *(*= a set).
b. *You can read a book, make a drawing* ***or something*** (= a class).
c. *You order, wait for food, urge the waiter because you are hungry, then wait again* ***and so on*** (= a frame).

[12] According to Mauri & Sansò (2020), languaging is “the process of making meaning and shaping knowledge and experience through language […]. *Languaging* thus refers to the activity performed in speech, which is an ongoing process constantly evolving and developing”. French linguists make use of the more or less corresponding NP ‘activité langagière’, which underlines the dynamic process (see, for instance, Bronckart 2007). Recurrent discourse patterns in the languaging activity may lead to stereotyped forms in a constructionalization process, as might be the case of *I dunno* from *I don’t know,* or French [ʃəˈpa] from *je (ne) sais pas,* used as general extenders marking indefiniteness at the end of a list (see Traugott & Trousdale 2013: 20, who speak of constructionalization as the creation of a $form_{new}$–$meaning_{new}$ pairing).

The three sentences do contain AHCs, though they represent different situations. This means that AHCs can be created via a large set of sentence types, and are not bound to a particular syntactic structure.

van der Auwera & Sahoo (2020: ex. (2a)) say that in a sentence like

(10) *I want such a cat.*

the function of *such* is to create an AHC during the discourse: contrary to the indefiniteness of the object alluded to in (6), the wanted cat is a definite exemplar of a newly created category, namely, CATS ENDOWED WITH THIS AND THAT PROPERTIES. Accordingly, it is quite possible that there are no limits to category building, provided that appropriate cues, like *such,* mark the sentence as an AHC.[13]

A second, crucial distinction obtains between *activity* and *category*: in the section "Lexicalization of goal-derived categories" in his 2021 article Barsalou (2021: 57) states that there is a surprising number of goal-derived categories that are lexicalized and he considers "the activity of *eating* and lexicalizations of categories associated with its important semantic roles, such as *diner* (agent), *food* (object), *utensil* (instrument), *eatery* (location), and *breakfast* (time)". Simple nouns such as *cat* or activities such as 'eating' can potentially be capable, via a bottom-up procedure, of opening the way to a (natural) category FOOD, composed of *hamburger*, *sandwich*, *egg*, *bread*, *salmon,* ***and so on***. In turn, *hamburger* may be considered as a member of the sub-category BURGER, together with *cheeseburger*, *fishburger*, *veganburger,* ***and so on***. Conversely, also Rosch's "natural category" BIRDS could also be a subcategory of ANIMALS and ANIMALS a subcategory of LIVING BEINGS. The risk of an endless (sub-)categorizing process is evident. Paradoxically, this seems to be in keeping with Smith & Samuelson's thesis (1997) that all categories are *ad hoc* and natural taxonomic categories like BIRDS, HUMANS, etc., do not exist and their lexicalization can be very arbitrary and different according to different cultures. Casasanto & Lupyan (2015) argued that there are in fact no stable categories that would be entrenched ready-made in people's minds: all categories emerge from current situations since people create them on the fly (see Moravcsik 2020). Consequently, one

[13] A distinction which is not always observed has to be kept in mind: namely, the distinction between linguistic tools introducing/concluding an AHC and the AHC in itself. It is not appropriate to write that "French *tel* or English *such* […] are essentially one-member categories" (van der Auwera & Sahoo 2020: conclusion): *tel* and *such* are linguistic tools capable to introduce/signal categories (and even one-member categories), but *per se* they are not a category.

could conclude that the very concept of category is useless, a conclusion which seems very counterintuitive, if we consider what we know about cognitive psychology and cognitive strategies such as making mind maps, association, mnemonics, etc.

5. Conclusion

As discussed above, the boundaries of a category and of an AHC may often be rather fuzzy in the sense of Wittgenstein's *Familienähnlichkeit* (family resemblance). Chauveau-Thoumelin (2018: 186) maintains that "a category with fuzzy boundaries […] is context depending". The more the category X is vague and undetermined, as is particularly the case with AHCs, the more examples are necessary to define the category by general extenders (see ex. (1), list constructions (ex. (9a)), lexemes such as *genre, problem, question,* defined by Chauveau-Thoumelin (p. 191) as "shell nouns":

(11) C*'est pour un roman historique,* ***genre*** *Dumas*
'It's for a Dumas-**like** historical novel.' (Chauveau-Thoumelin, 2018, ex. (3))

However, as we have seen in the previous sections, AHCs with just one example as in (6) are not frequent and even "shell nouns" like *genre, tipo* offer many instances with more than one example:[14]

(12) *Il existe de tout petits bacs de 250 ml avec de nouveaux parfums* ***genre*** *bergamotte, marron glacé, spéculoos, absinthe, chocolat blanc.*
'There are tiny, 250 ml containers with new gourmet flavours like bergamot, marron glacé, speculoos, absinth, white chocolate.' (Chaveau-Thoumelin, 2018: 183)

We may conclude that the classical, traditional definition of category, as reported above (section 1), does not apply to the AHCs. If we accept the rigid definition given in the dictionaries, then we should find a different name for the AHC, e.g., "ad hoc ensemble", or "ad hoc group". However, the term "ad hoc category" has already

[14] An anonymous reviewer notes that a quantitative study would be needed. This is correct from a theoretical viewpoint. However, given the unlimited possibility of new AHCs and the absence of a dedicated corpus, it is practically unfeasible. As I said at the beginning of this paper, my examples are drawn from the extant literature.

acquired a respectable citizenship among linguists and so we will go on speaking of AHCs. The aim of my reflections, discussing the recent literature, has simply been to observe that "ad hoc categories" are a very particular type of "category", a construct endowed with its particular rules.[15]

Abbreviations

ADJ = adjective
GEN = genitive
ECHO = echo-word construction
M = masculine
NEG = negation
NOM = nominative
PAST = past
POL = polite register
PRES = present

References

Arcodia, Giorgio F. & Caterina Mauri. 2020. Exemplar-based compounds: The case of Chinese. *Language Sciences* 81. 1-21.

Barotto, Alessandra. 2018. The role of exemplification in the construction of categories: the case of Japanese. *Folia Linguistica Historica* 39. 37-68.

Barsalou, Lawrence W. 1983. Ad hoc categories. *Memory and Cognition* 11. 211-227.

Barsalou, Lawrence W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22. 577-660.

Barsalou, Lawrence W. 2010. Ad hoc categories. In Patrick C. Hogan (ed.), *The Cambridge Encyclopedia of the Language Sciences*, 87-88. Cambridge: Cambridge University Press.

Barsalou, Lawrence W. 2021. Categories at the interface of cognition and action. In Ilaria Fiorentini, Caterina Mauri & Eugenio Goria (eds.), *Building Categories in Interaction: Linguistic Resources at Work*, 35-72. Amsterdam: John Benjamins Publishing Company.

[15] The following sentence I have found in the DWDS (see fn. 2) is telling enough about the terminological difficulties: "*Diese kollektive Identität bestimmt den **Kreis** derer, die sich als Angehörige derselben sozialen **Gruppe** verstehen und von sich unter der **Kategorie** der ersten Person Plural sprechen können".* 'This collective identity defines the **circle** of those people who consider themselves as belonging to the same social **group** and may speak about themselves using the **category** of the first plural person' (my emphasis, Habermas 1981: 95). Habermas uses here 'category' in the sense linguists refer to morphological distinctions while 'group' in such a context means 'category'. It is a good example of how fuzzy definitions can be.

Bronckart, Jean-Paul. 2007. L'activité langagière, la langue et le signe, comme organisateurs du développement humain. *Langage et société* 121-122. 57-68.

Casasanto, Daniel & Gary Lupyan. 2015. All Concepts are Ad Hoc Concepts. In Eric Margolis & Stephen Laurence (eds.), *The Conceptual Mind: New directions in the study of concepts*, 543-566. Boston: The MIT Press.

Chauveau-Thoumelin, Pierre. 2018. Exemplification and *ad hoc* categorization: The *genre*-construction in French. *Folia Linguistica Historica* 39. 177-199.

De Mauro, Tullio. 2000. *Il dizionario della lingua italiana.* Torino: Paravia.

Diccionario de la lengua Española. «Categoría». Real Academia Española. (https://dle.rae.es/categor%C3%ADa?m=form).

Digitales Wörterbuch der deutschen Sprache. «Gruppe». Berlin-Brandenburgische Akademie der Wissenschaften. (https://www.dwds.de/wb/Gruppe).

Goria, Eugenio & Francesca Masini. 2021. Category-building lists between grammar and interaction. In Caterina Mauri, Ilaria Fiorentini & Eugenio Goria (eds.), *Building Categories in Interaction: Linguistic resources at work*, 73-110. Amsterdam: John Benjamins Publishing Company.

Haspelmath, Martin. 1993. *A Grammar of Lezgian*. Berlin-Boston: De Gruyter Mouton.

Habermas, Jürgen. 1981. *Theorie des kommunikativen Handelns* - Bd. 2. *Zur Kritik der funktionalistischen Vernunft*. Frankfurt: Suhrkamp.

Inglese, Guglielmo & Ulrich Geupel. 2018. The encoding of *ad hoc* categories in Sanskrit: A synchronic and diachronic analysis of “compounds” with *ādi-*. *Folia Linguistica Historica* 39. 225-252.

Kallergi, Haritini. 2015. *Reduplication at the word level. The Greek facts in typological perspective.* Berlin-Boston: De Gruyter Mouton.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar,* Volume 1*: Theoretical Prerequisites.* Stanford: Stanford University Press.

Larousse, Dictionnaire français monolingue. «Espèce». Larousse. (https://www.larousse.fr/dictionnaires/francais/esp%C3%A8ce/31030).

Lo Baido, Maria Cristina. 2018. Categorization *via* exemplification: evidence from Italian. *Folia Linguistica Historica* 39. 69-95.

Lo Baido, Maria Cristina. 2023. Tra modalità e categorizzazione indessicale: il caso di *sapiddu* e *chi sacciu*. *Cuadernos de Filología Italiana* 30. 135-161.

Magni, Elisabetta. 2018. Collective suffixes and ad hoc categories: from Latin *–ālia* to Italian *–aglia*. *Folia Linguistica Historica* 39. 201-224.

Masini, Francesca, Caterina Mauri & Paola Pietrandrea. 2018. List constructions: Towards a unified account. *Italian Journal of Linguistics* 30. 49-94.

Mauri, Caterina. 2017. Building and interpreting ad hoc categories. In Joanna Blachowiak, Cristina Grisot, Stephanie Durrleman-Tame & Cristopher Laenzlinger (eds.), *Formal Models in the Study of Language*. A Festschrift for Jacques Moeschler, 297-326. Berlin: Springer.

Mauri, Caterina. 2021. Ad hoc categorization in linguistic interaction. In Caterina Mauri, Ilaria Fiorentini & Eugenio Goria (eds.), *Building Categories in Interaction: Linguistic resources at work*, 9-34. Amsterdam: John Benjamins Publishing Company.

Mauri, Caterina & Anna Giacalone Ramat. 2015. *Piuttosto che*: dalla preferenza all'esemplificazione di alternative. *Cuadernos de Filología Italiana* 22. 49-72.

Mauri, Caterina & Andrea Sansò. 2018a. Un approccio tipologico ai '*general extenders*', in Marina Chini & Pierluigi Cuzzolin (eds.), *Tipologia, acquisizione, grammaticalizzazione*, 63-72. Milano: FrancoAngeli.

Mauri, Caterina & Andrea Sansò. 2018b. Linguistic strategies for ad hoc categorization: theoretical assessment and cross-linguistic variation. *Folia Linguistica Historica* 39. 1-35.

Mauri, Caterina & Andrea Sansò. 2020. Ad hoc categorization and *languaging*: the online construction of categories in discourse. *Language Sciences* 81. 1-7.

Mauri, Caterina, Ilaria Fiorentini & Eugenio Goria. 2021. *Building Categories in Interaction: Linguistic resources at work*. Amsterdam: John Benjamins Publishing Company.

Mihatsch, Wiltrud. 2018. From ad hoc category to ad hoc categorization: The proceduralization of Argentinian Spanish *tipo*. *Folia Linguistica Historica* 39. 147-176.

Moravcsik, Edith. 2020. The place of ad hoc categories within the typology of plural expressions. *Language Sciences* 81.

Oxford English Dictionary. «Group». Oxford University Press. (https://www.oed.com/dictionary/group_n?tab=factsheet#2559543)

Rosch, Eleanor. 1973. Natural categories. *Cognitive Psychology* 4. 328-350.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson & Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8. 382-439.

Sammarco, Carmela. 2021. Online text mapping. The contribution of verbless construction in spoken Italian and French. In Caterina Mauri, Ilaria Fiorentini & Eugenio Goria (eds.), *Building Categories in Interaction: Linguistic resources at work*, 211-238. Amsterdam: John Benjamins Publishing Company.

Schneider, Erich. 1965. *Theorie des Wirtschaftskreislaufes*. Tübingen: Mohr.

Smith, Linda B. & Larissa K. Samuelson. 1997. Perceiving and remembering: Category stability, variability and development. In Lamberts Koen & David Shanks (eds.), *Knowledge, Concepts and Categories*, 161-195. Hove: Psychology Press.

Stolz, Thomas. 2003/04. A new mediterraneanism: Word iteration in an areal perspective. *Mediterranean Language Review* 15. 48-62.

Stolz, Thomas. 2018. (Non-)Canonical reduplication. In Aina Urdze (ed.), *Non-Prototypical Reduplication*: Studia Typologica 22, 201-277. Berlin-Boston: De Gruyter Mouton.

Traugott, Elizabeth & Graeme Trousdale. 2013. *Constructionalization and Constructional Changes.* Oxford: Oxford University Press.

Treccani Vocabolario. 2003. «Categoria». Istituto dell'enciclopedia italiana. (https://www.treccani.it/vocabolario/categoria/)

van der Auwera, Johan & Kalyanamalini Sahoo. 2020. Such similatives: a cross-linguistic reconnaissance. *Language Sciences* 81.

Voghera, Miriam. 2012. Chitarre, violino, banjo e cose del genere. In Miriam Voghera & Anna Maria Thornton (eds.), *Per Tullio De Mauro. Studi offerti dalle allieve in occasione del suo 80° compleanno*, 341-364. Roma: Aracne.

Voghera, Miriam. 2013a. A case study on the relationship between grammatical change and synchronic variation: the emergence of *tipo[-N]* in Italian. In Anna Giacalone Ramat, Caterina Mauri & Piera Molinelli (eds.), *Synchrony and diachrony: a dynamic interface*, 283-312. Amsterdam: John Benjamins Publishing Company.

Voghera, Miriam. 2013b. *Tipi di* tipo *nel parlato e nello scritto*. In Immacolata Tempesta & Massimo Vedovelli (eds.), *Di Linguistica e di Sociolinguistica. Studi offerti a Norbert Dittmar*, 185-195. Roma: Bulzoni.

CONTACT

paoram@unipv.it