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.

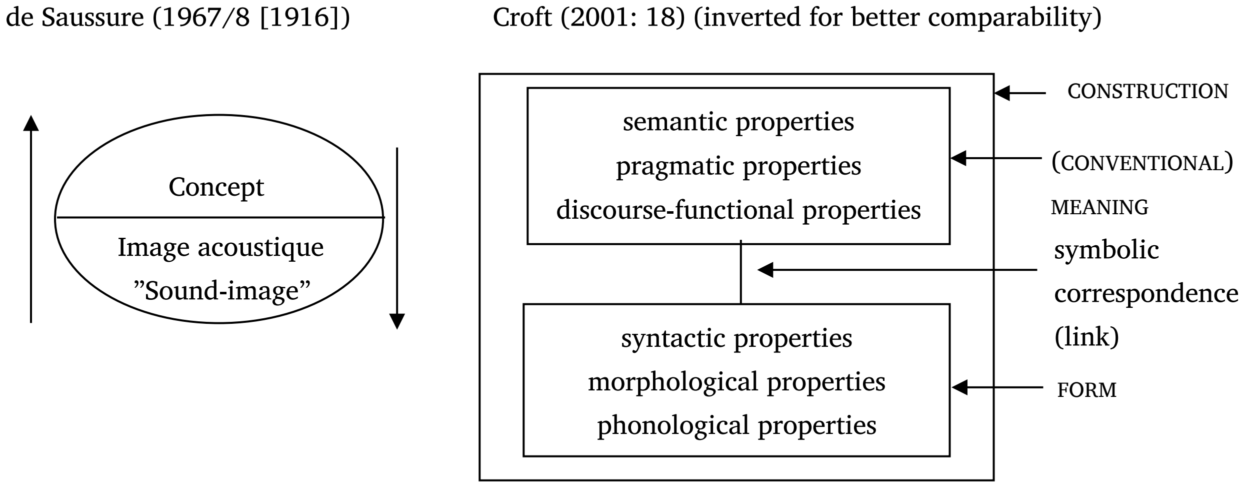

Figure 1: Symbolic link between marker and meaning in structuralism and in construction grammar.

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.

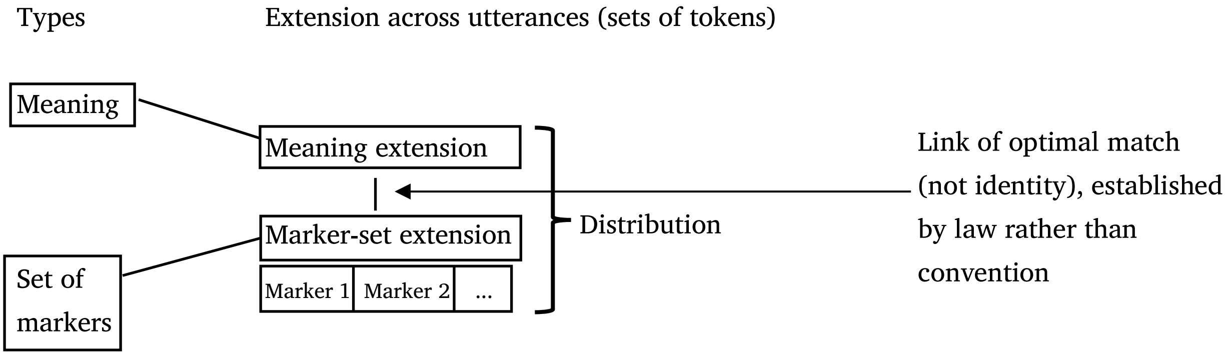
Types
Extension across utterances (sets of tokens)
Meaning
Meaning extension
Set of markers
Marker-set extension
Marker 1
Marker 2
...
Distribution
Link of optimal match (not identity), established by law rather than convention

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 -2logλ 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
1	1421.5	m	4	conn	244
2	1405.7	m	5	#conn	229
3	975.3	m	6	#conna	172
4	974.2	m	5	conna	179
5	815.5	m	4	#sav	140
6	815.5	m	3	sav	140
7	620.0	m	4	onna	183
8	610.4	m	5	#sach	83
9	607.5	m	3	nna	183
10	596.2	m	4	sach	83
…					
47	368.0	w	6	#sais#	48
…					
11717 candidates in total					

Columns

A: Rank,

B: Collocation value (Dunning's log-likelihood),

C: Candidate type (w = word-form, b = bigram, m = morph),

D: Number of characters of the candidate (word boundary is counted as a character; if the collocation value is the same, longer candidates are ranked higher)

E: Candidate (# stands for word boundary),

F: Number of occurrences within search distribution

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

A	B	C	D	E	F	G
1	1421.5	m	conn	244	331	1421.5
5	815.5	m	#sav	375	504	2369.5
8	610.4	m	#sach	455	593	3148.7
40	368.2	w	sais	493	635	3573.6
80	174.4	w	sait	509	651	3782.1
330	16.5	w	sût	513	655	3837.8

Columns

A: Rank,

B: Candidate collocation value (Dunning's log-likelihood),

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 (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.

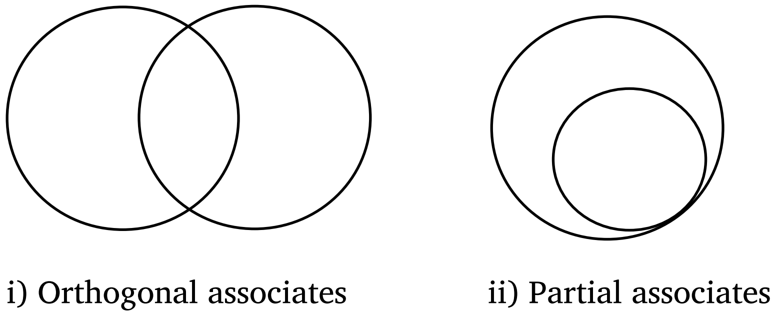

Figure 3. Illustration of orthogonal and partial associates.

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*	[#chɔ̃́ŋ#]	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.

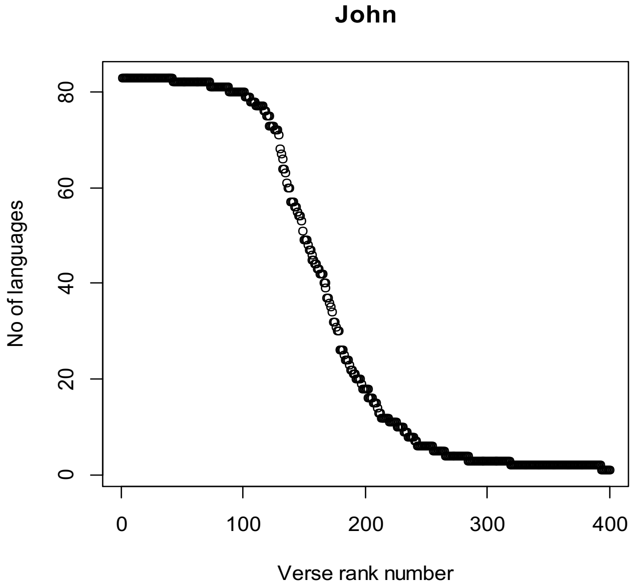

Figure 4: Occurrences of markers for 'John'.

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) *juandicue* (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) *jíiꞌ˜ 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 <nɨ-> 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 *l̲ama* [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.

<table>
<tr><th>Language</th><th>Extracted markers</th><th>Verses</th><th>Coverage</th><th>Dedication</th><th>Coll. value</th></tr>
<tr><td>Mandarin Chinese**</td><td>[i1dao* | ao3d | #ren4shi# | #qi3bu4 zhi1# | #ren4 de2# | #ren4 chu1#]</td><td>472</td><td>88.06%</td><td>82.95%</td><td>3682.07</td></tr>
<tr><td>Zarma**</td><td>[#bay#]</td><td>468</td><td>87.31%</td><td>83.13%</td><td>3651.21</td></tr>
<tr><td>Pennsylvania German</td><td>[#viss | #vays | #gvist# | #gekend# | #ich kenn# | #eisicht# | #unbekand]</td><td>450</td><td>83.96%</td><td>83.33%</td><td>3479.26</td></tr>
<tr><td>...</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Southern Nambikuára**</td><td>[a3la3kx | ko̱3nh | e3wxe]</td><td>209</td><td>38.99%</td><td>40.98%</td><td>643.28</td></tr>
<tr><td>Kalam</td><td>[niŋb | #niŋr]</td><td>323</td><td>60.26%</td><td>23.30%</td><td>596.07</td></tr>
<tr><td>Fasu</td><td>[#hemaka | himete | #asera]</td><td>442</td><td>82.46%</td><td>14.29%</td><td>490.98</td></tr>
</table>

* *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	Coverage of	Dedication of	Colloc.
		covered	set	set	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# | #аса#]	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