Linguistic Typology at the Crossroads 5-2 (2025): 66-104 ISSN 2785-0943

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

NICHOLAS EVANS[1,2], DANIELLE BARTH[1,2], WAYAN ARKA[1,2,3], HENRIK BERGQVIST[4], CHRISTIAN DÖHLER[5], SONJA GIPPER[6], YUKINORI KIMOTO[7], DOMINIQUE KNUCHEL[8], DANIEL MAJCHRZAK[9], HITOMI ŌNO[10], EKA PRATIWI[11], SASKIA VAN PUTTEN[12], ANDREA C. SCHALLEY[13], ASAKO SHIOHARA[14], STEFAN SCHNELL[15], YANTI[16]

[1]School of Culture, History & Language, Australian National University; [2]Centre of Excellence for the Dynamics of Language, Australian National University; [3]Universitas Udayana; [4]Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg; [5]Berlin-Brandenburg Academy of Sciences and Humanities; [6]Department of Linguistics, University of Cologne; [7]Graduate School of Humanities, Osaka University; [8]Department of Linguistics, University of Bern; [9]School of Literature, Languages and Linguistics, Australian National University; [10]Reitaku University; [11]Faculty of Foreign Languages, Universitas Mahasaraswati Denpasar; [12]Centre for Language Studies, Radboud University Nijmegen; [13]Department of Language, Literature and Intercultural Studies, Karlstad University; [14]Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies; [15] Institute for the Interdisciplinary Study of Language Evolution, University of Zurich; [16]Atma Jaya Catholic University of Indonesia

Submitted: 06/02/2025 Revised version: 27/11/2025
Accepted: 27/11/2025 Published: 25/02/2026

Abstract

This article examines the question of whether complementation structures are cross-linguistically universal by using two different cross-linguistic corpora, each drawing on the same thirteen languages, spanning every continent. One is SCOPIC, the *Social Cognition Parallax Interview Corpus*, specifically designed to elicit material rich in grammatical categories relevant to social cognition; for each language in our sample this was balanced by

[*] Note (added on 3 March 2026): This article was updated to include Stefan Schnell in the list of authors. See the Corrigendum published at https://doi.org/10.60923/issn.2785-0943/24308.

a "general corpus" of roughly the same size with no specific targeting of domains. We find that, while complementation is widespread, it is not universal within the languages in our sample: in some it is absent entirely and in others it is extremely rare. Of the structural alternatives used to achieve the same functional goal by far the commonest is quoted speech, suggesting that in the evolution of linguistic structures it is heteroglossia, the embedding of one person's words in another's, that is a more basic phenomenon, from which complementation structures then evolve in many but not all languages.

Keywords: complementation; clause combining; corpus-based typology; clustering; parallax corpora; quoted speech; propositional framing

1. Introduction

In this article we ask two questions. The first is typological: are complementation constructions universal? This is a debate of long standing within typology (e.g. Aikhenvald & Dixon 2006[1]; Klamer 2000) but what is new about our study is that we subject the question to close cross-linguistic scrutiny using parallax corpus techniques. The second is methodological: does the SCOPIC corpus, a parallax corpus designed to elicit enriched naturalistic data on the language of social cognition (Barth & Evans 2017a, 2017b) provide a representative sample of language use more generally when compared against other datasets?

Recursion, of which complementation constructions are one of the three major types alongside relative clauses and adnominal recursion, is a cornerstone of much linguistic theorising, including accounts of generativity in grammar, vital steps in language evolution, and the development of *theory of mind* in social cognition, topics we summarise in (§2). Against this background, we need to ask a more basic empirical

[1] For example, in the abstract of their monograph on complementation Aikhenvald & Dixon (2006) write: "A **complement** clause is used instead of a noun phrase; for example one can say either I heard [the result] or I heard [that England beat France]. Languages differ in the grammatical properties of **complement** clauses and the types of verbs which take them. Some languages lack a **complement** clause construction but instead employ other construction types to achieve similar ends; these are called complementation strategies" (https://academic.oup.com/book/51146; accessed 2025-12-28), and their cross-linguistically inclusive collection gives examples of languages lacking complementation constructions. However, we believe that using the term *complementation strategies* continues to introduce a structural bias into the discussion, which is why we will use the more neutral term *propositional framing* as a functional label – see §3.3.

question: how widespread do complement constructions turn out to be, in fact, when we look at corpora of actual language use? Do all languages have them? Are they used more often for some types of complement than for others?[2] What are the functional alternatives to complement constructions? This is the clutch of typological questions we set out to answer in this paper.

In answering these questions, we draw on two distinct types of data (§3). One is the Family Problems picture task (San Roque et al. 2012) and the SCOPIC parallax corpus resulting from it (Barth & Evans 2017a, 2017b). Within the SCOPIC corpus (Barth & Evans 2024) we now have data for over 30 languages from every continent. In Kimoto et al. (2024), we drew on fourteen of these and examined what constructional alternatives to complement constructions are frequently attested in the SCOPIC data. However, it is important to ask how far it is representative of a broad sample of natural language use. In this paper, therefore, for each language represented, we also use a comparative "free" dataset – assembled from one or more corpora produced for other purposes; in §3.2 we outline how these free datasets, roughly comparable in size for each language to the corresponding SCOPIC sub-corpus, were built up across our language sample. Though there is substantial overlap in the languages used in the two studies, Kimoto et al. (2024) included Ilokano, Ku Waru, Jinghpaw, Japanese and Sibe, not in the present study, while this study includes Balinese, Avatime, Kogi, Komnzo, Vera'a and Yurakaré, not in the Kimoto et al. study. This reflected the availability of resources and time for the language-expert colleagues in our team to compile the "free" datasets needed for this part of the study.

In §3.3 we turn to the annotation schemata used in our analysis. Both sets of corpora were annotated for complement constructions and their functional equivalents, following an annotation scheme standardised across SCOPIC investigators; this uses a subset of the much larger set of conventions for annotating the SCOPIC corpus set out in Barth et al. (2024). These construction-type labels are applied to all propositional framing constructions in the corpus – basically all constructions which can frame propositions in the way that is done by complement constructions in at least some languages, as set out in the seminal article by Noonan (1985). We use this set of functional definitions to annotate structural alternatives to

[2] Again, we are by no means the first to ask this question. See e.g. Spronck & Casartelli (2021) for a wide-ranging survey of how complementation is likely to have originated through the intextualisation of direct speech; the extensive literature on how complementisers get recruited from 'say' verbs (Frajzyngier 1984, Klamer 2000, Saito 2021) implies that it is represented-speech complements which appear first in the grammaticalisation of complement constructions.

complementation in expressing comparable functions across the languages of our sample, in other words to examine the cross-linguistic semasiological possibilities for each relevant onomasiological choice. Thus, alongside complement constructions we also include such alternatives as paratactic quoted speech, inflections (e.g. desiderative inflections instead of 'want' verbs) and particles (e.g. counterfactual particles instead of verbs like *believe*). We likewise develop a set of functions (e.g. *utterance*, *thought*, *fear*) based on the functional range of complement-taking predicates in Noonan (1985).

In §4 we go back to our SCOPIC corpus and examine how far languages adopt the same coding strategies to express propositional framing. As we will show, the proportion of propositionally framed constructions that use complement constructions is highly variable cross-linguistically. While complement constructions are used commonly in some languages there are others in which they appear very rarely or not at all in our corpus, their place being taken by other constructions like paratactically linked direct quotation.

In §5 we evaluate the representativeness of the SCOPIC corpus by comparing the results for each language-specific SCOPIC sub-corpus with a comparably sized dataset from the same language. Overall, the results are remarkably similar, but there are nonetheless differences for some of the languages in our sample, which we discuss in this section.

In §6 we conclude by furnishing answers to the two questions posed at the outset. Taking these in reverse order, and beginning with the methodological question, there is broad agreement between our two types of data. Once we factor out specific reasons altering the occurrence of some strategies in the sub-corpora, we can take the SCOPIC corpus as an ecologically valid representation of overall frequencies of complementation or its functional equivalents. The great advantage of the SCOPIC data is that it permits close alignment of described scenarios across a wide range of languages while not biasing participant's expressive decisions through the effects of a "founder" language, such as one finds in parallel corpora (e.g. Mayer & Cysouw 2014), where one language is the source[3] and the others are translations from that source. In other words, it does not carry forward particular decisions about what to encode grammatically in an original "founder" language, from which translations are produced; rather speakers and signers of each language make their own direct decisions about what to encode based on the stimulus set.

[3] For the full Bible, of course, the source language is sometimes Biblical Hebrew (most of Old Testament), sometimes Aramaic (portions of Daniel and Ezra), and Greek for the New Testament. This does not alter the fact that, for whatever portion of the Bible we are talking about, one language is a "founder" or ultimate source for translations into all other languages.

An affirmative answer to the methodological question then gives us confidence to tackle the first, theoretical question. We find that complementation is not a universal encoding strategy, and some languages draw on other structures to achieve the same expressive ends, in particular by using direct quotation. The intertextuality afforded by direct quotation, we conclude, is a more cross-linguistically robust source of what may then, optionally and in only some historical trajectories, evolve into syntactically recursive complement constructions.

2. The claimed centrality of complementation to syntactic theory and social cognition

Over the last six decades, complementation has been central to arguments about the recursiveness of language, about the key steps in the evolution of language proper from earlier communicative systems and about the development of social cognition. Consider the multiple levels of complementation exhibited in (1):

(1) *All readers of this journal suspect that [Caterina thinks that [many typologists don't believe that [convincing typologies need naturally occurring discourse data]]]].*

The rewrite rules that sentences like (1) are believed to need (of the type (i) S → … NP; (ii) NP → S) were used by Chomsky (1957) and subsequent developments of generative grammar to argue for the indispensability of powerful transformational grammars necessary to account for the generativity of language. Of course, complement structures are just one type of syntactic recursion, alongside relative clauses and adnominal modification structures. Thus, we can represent the visual recursion shown in Figure 1 by a complementation structure, in (2), a recursive relative clause structure, as in (3), and the recursion of adnominal NPs, as in (4).

(2) *Chomsky and Halle remember that [back in the main photo they were thinking that [they were dressed rather formally in the river scene]]]].*

(3) *The photo [that shows a scene [where Chomsky and Halle hold a picture [that portrays them as young students]]] was taken by Michael Yoshitaka Erlewine.*

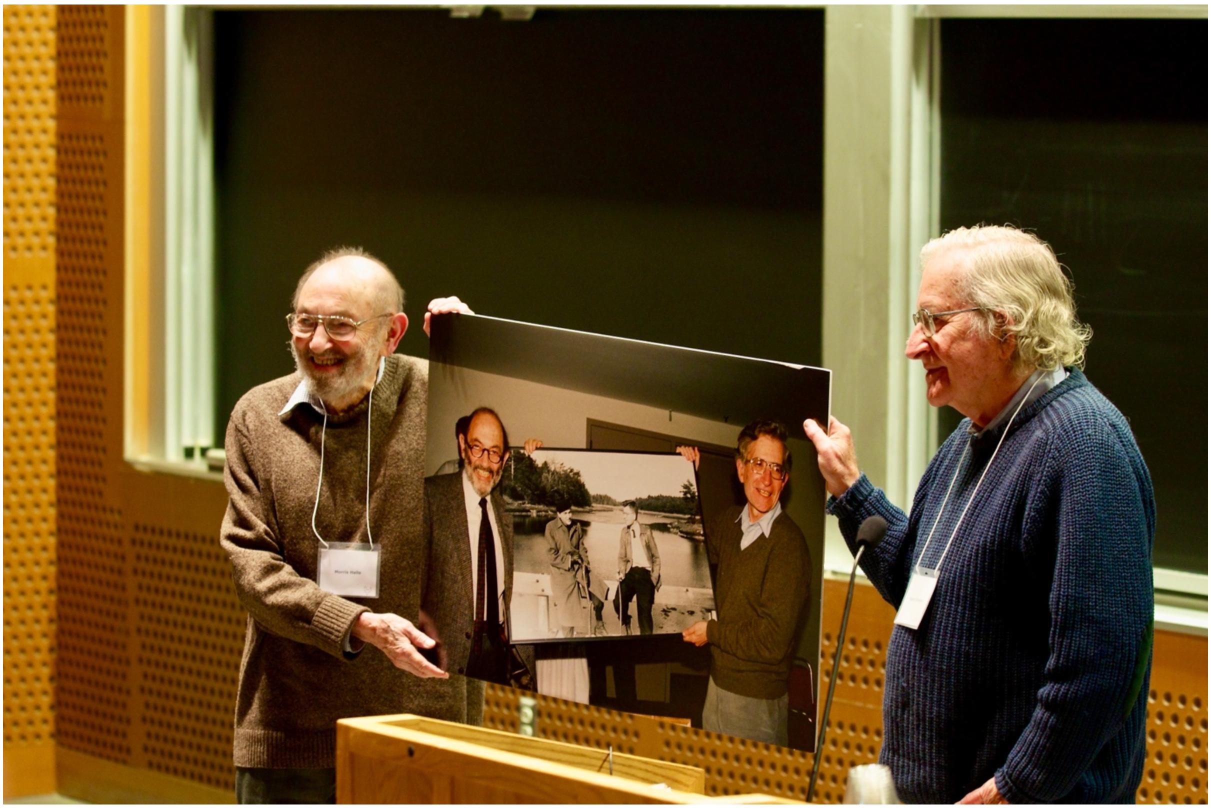

Figure 1. At Ling50@MIT, Morris Halle and Noam Chomsky holding a 1988 picture of them holding a picture of them in 1953[4].

(4) *We here reproduce [a slide of [an image of [a photo of [a scene of [Chomsky and Halle's youthful days]]]].*

Among these three types of syntactic recursion, complementation has a privileged status in important neighbouring fields, such as the psychology of social cognition. Psychologists of social cognition such as De Villiers (2000) and De Villiers & De Villiers (2003, 2014) argued that the acquisition of *theory of mind* and mastery of *false belief tasks* in developing children went hand-in-hand with the acquisition of complement constructions that allowed them to represent states of affairs as being held in the minds of particular social actors. De Villiers & Pyers (2002: 1057) discuss

[4] Photo and concept credit. The idea for this photo came from Sabine Iatridou (watch https://www.youtube.com/watch?v=csE-MsT_NN0 from around 1:40:40; accessed 2025.12.28) and it is discussed in the following blog by Kai von Fintel: https://www.kaivonfintel.org/morris-noam-recursion/ (accessed 2025.12.28) and the photo from Michael Yoshitaka ("Mitcho") Erlewine, with whose kind permission it is reproduced here. Our thanks to all three for their kind help and permission with this photo and story.

the claim that the development of complement structures in children is the best predictor of their performance on false-belief tasks: "[the] effect here is likely to be bi-directional, namely, understanding mental states undoubtedly makes it more probable and easier for the child to encode events in terms of people's beliefs, motives, intentions, emotions and so forth". The rationale for this argument is that mental attitude predicates, like those which are multiply embedded in (1), are the way that humans represent the mental states of others, and that this is done using conceptual analogues of the recursive complement structures shown here. And indeed, various language-specific modifications of complement structures are widespread cross-linguistically. For example, (5) from Japanese, produced spontaneously during one run of our SCOPIC task (§3.1), is an example with three levels of embedding of complement clauses. As is typical for a left-branching language like Japanese the embedded complement clauses each precede their complement-taking predicate.

(5) Japanese (Kazuya Inagaki - SocCog-jpn01-ikst3.eaf - 05:47.9-5:53.1)

de	*[[[de-ta*	*ato*	*wa*	*konna*	*koto*	*ga*
then	go.out-PST	after	TOP	such	thing	NOM

mat-teru	*daroo-na*	*to]*	*iu*	*no*	*o*
wait-PROG	may-FP	QUOT	say	COMP	ACC

soozoo-si-teiru]	*zu*	*da*	*to]*	*omoi-masu.*
guess-do-PROG	picture	COP	QUOT	think-POL

'Then, (I) think [(this) picture shows that [(he) is imagining [what will happen to (him) after going out (of prison)]]].'

Nonetheless, these structural solutions are not universal, and they are only used in a subset of the world's languages: languages have many other methods for showing mental attitudes. For example, in the Australian language Dalabon, the particle *djehneng* or its variant *yangdjehneng,* roughly 'believedly', is used to cast a statement as someone's belief rather than an actual fact; the identity of the believer is established pragmatically, e.g. through mention of a relevant belief-holding candidate in a preceding clause. Consider the following example from a text about a songman J's anger when he believes that another singer is singing J's compositions and passing them off as his own. Translation (a) puts this into more normal (complementiser-using) English, using the verb *think* and an embedded complement, while (b) translates this more literally using an adverb to render the syntactic effect of *yangdjehneng:*

(6) Dalabon (Dal20090624.musdisc.mtandothers.eaf 2:58.4-3:01.8[5])

ka-h-kangurdinjirrmi-nj
3SG.SBJ-R-get.angry-PST.PFV

yangdjehneng *bûrra-h-marnû-dulu-djirdm-ey*
believedly 3DU.A > 3SG.OBJ-R-BEN-song-steal-PST.PFV

kanh *kodj-no* *barra-h-wayirni-nj*
DEM tune-3SG.POSSD 3DU.SBJ-R-sing-PST.IPFV

a. 'He got upset (that)/(because he thought that) the two of them had stolen it and were singing his song.'
b. 'He got upset; believedly they two had stolen it and were singing his song.'

Another Dalabon method for expressing mental attitudes without using complementation structures is shown in (7), this time using the adverbial prefix *molkkunh-* inside a polysynthetic verb.

(7) Dalabon (Evans et al. 2004: 245)

[Context: NE and a friend had turned up at the speaker's community the night before, without having been able to let her know, for want of a telephone, and camped nearby rather than imposing on her. Next day she reproached us:]

*de-h-**molkkunh**-bo-ng* *dabangh* *nahda,*
2DIS.SBJ-R-unbeknownst-go-PST.PFV yesterday hither

mak *yila-bengkey.*
NEG 1PL.SBJ > 1PL.OBJ.IRR-know-IRR

Lit. 'The two of you came here yesterday, unbeknownst, we didn't know.'
More English-like translation: 'We didn't know that the two of you had come here yesterday.'

[5] Available at https://www.gerlingo.com/language_detail.php?langID=7 (accessed 2025.12.18).

As with *yangdjehneng*, which in many ways is its semantic obverse (*yangdjehneng* specifying someone's belief, *molkkunh-* someone's ignorance), the clause containing it does not overtly specify the holder of the mental attitude. But as can be seen from these two examples, it is common (though not necessary) for this mental-attitude-holder to be specified in another clause nearby, such as *kahkangurdinjirrminj* 'he got angry' in (6) or *mak yilabengkey* 'we didn't know' in (7). In each of these two cases the identity of the mental-attitude-holder is not specified by the attitude-projection particle or affix itself (*yangdjehneng* 'believedly', *molkkunh-* 'unbeknownst') but found on another verb, whose subject identifies the holder of the mental attitude, in a paratactically juxtaposed clause. See Evans (2021) for many more examples, including cases where there is no neighbouring clause to specify the mental attitude-holder and this is left sheerly to pragmatic inference.

Here, then, Dalabon shows how it is quite possible for a language to represent mental predicates without the use of syntactic complement structures, in this case by a propositional-attitude particle showing projected belief. We will see later that it has other means as well, through the use of quoted speech; see also Kimoto et al. (2024) for various other constructional alternatives found in the SCOPIC corpora.

Within the field of language development, there have also been researchers citing other means of representing mental attitudes without the use of complementation strategies. For example, Matsui et al. (2009) in their comparison of theory-of-mind development in German and Japanese three-year-olds showed superior performance of Japanese as compared to German children on Theory of Mind tasks. They impute this to the high-frequency use of the illocutionary particle *yo* in Japanese, which signals a person's belief about what they are saying, as opposed to alternatives like *kana* 'maybe', and hence cues children early to attend to what those around them believe to be the case. Even though both Japanese and German have complement constructions in frequent use, for Japanese children it appears to be the illocutionary particles rather than the complement structures which are doing the heavy lifting in terms of scaffolding emergent social cognition, at least in the realm of theory of mind.

Considerations like these mean that, rather than assuming that complementation is cross-linguistically universal, and the sole structural means of representing mental attitude predicates, we should approach the question empirically and cross-linguistically. The approach we take is to define the relevant phenomena functionally and then determine what structure types are used to represent them using cross-linguistic corpora.

3. Datasets and corpora: methods and annotation

We now turn to the corpora for gathering comparable data across our cross-linguistic sample and the methods we use to annotate them: the *Social Cognition Parallax Interview Corpus* (SCOPIC) (§3.1), specifically developed to elicit enriched data on social cognition, and the less targeted general corpora (§3.2) which can be used to evaluate the representativeness or otherwise of the SCOPIC corpus, and the annotation schema we use in both (§3.3).

3.1 The cross-linguistic SCOPIC corpus

The cross-linguistic SCOPIC is a *parallax* corpus, which we have defined elsewhere (Barth & Evans 2017b: 1) as involving "broadly comparable formulations resulting from a comparable task", to avoid the implications of *parallel* corpus that there will be exact semantic equivalence across languages. The rationale for a parallax approach is that, by giving each participant the opportunity to respond in their own way to a shared stimulus, we leave it up to them to express things the way they want, without the "founder bias" that comes from strict translation tasks. In translation tasks, the source language Is likely to nudge the translation to transfer certain semantic categories or certain structures to the target language.

At the same time, using a shared stimulus gives us good control[6] over the referential characteristics of the described event, allowing us to match formulations across languages and across individuals.

The stimulus set consists of 16 picture cards from the Family Problems Picture Task (San Roque et al. 2012), which was developed as a *broad spectrum task* as part of a project on the cross-linguistic grammar of social cognition, with the goal of eliciting a wide range of themes relevant to social cognition (kinship relations, expression of emotions, private predicates, social consequences of actions from benefaction to malefaction), but also thought, speech, fear, memory and wishes for the future. These latter categories are all, evidently, good candidates for deploying complementation structures – but the overall task was not designed with any specific investigation of complementation in mind.

[6] Though of course this control is not complete, since different people construe the stimulus pictures according to their own cultural and individual schemata – clothes given back to a prisoner emerging from gaol may be seen as his own, or as a new ranger uniform; individual characters may be seen as male or female and so forth.

Our task was designed so that participants would see the task as meaningful and engaging, empathise with the characters and situations, including strong emotional reactions at certain points (such as domestic violence), understand the graphic conventions (speech and thought bubbles), understand the task specifications (breakdown into subtasks; ordering of pictures; distribution between dialogic and monologic subtasks) and discuss freely, vividly and unselfconsciously. We hoped to balance two aims: on the one hand to elicit broadly comparable cross-linguistic data, and on the other to be a kind of cultural Rorschach blot that calls forth interesting culturally specific and grammatically specific elements.

The task structure allows people, working in pairs, the chance to construct their own stories, across four task stages (a) initial card-by-card description, (b) joint construction of a meaningful narrative sequence (this stage was designed to elicit vigorous discussion), (c) narration of the story to a third party who had not been present for stages (a) and (b) and therefore started with no common ground, (d) narration of the same story, in the first person, from the perspective of one of the characters. See San Roque et al. (2012) for a detailed description of the task.

Over the last decade we have been building the cross-linguistic SCOPIC corpus from around 30 languages of all continents, including one sign language (Auslan); see Barth & Evans (2017b) for a listing of 25 of these, though the set continues to grow, and Barth & Evans (2024) for the archived corpus data including transcriptions and translations. In the research on which this article was based, we draw on a subset of thirteen languages (Table 1).

The sub-corpora for individual languages range from 34:01 (Indonesian) to 6:30:31 (Balinese) with a total of 33:39:49 hours and a mean of 1:40:59; for fuller details of SCOPIC sub-corpora sizes see §3.2.

3.2 Supplementary comparison corpora

As described in §3.1, the Family Problems Picture Task was designed as a broad spectrum stimulus task for getting enriched data on the expression of categories relevant to social cognition. SCOPIC data often includes expressions of mental attitudes, desire, intention, emotion, reported quotation, etc. This could potentially lead to corpus bias effects, e.g. if the choice of stimulus set were to bias the proportion of mental attitude constructions.

Language	Family	Location
Arta [atz]	Austronesian	Philippines
Avatime [avn]	Niger-Congo	Ghana
Balinese [ban]	Austronesian	Indonesia
Dalabon [ngk]	Australian	Australia
English [eng]	Indo-European (Germanic)	Australia
German [deu]	Indo-European (Germanic)	Germany
G|ui [gwj]	Khoe-Kwadi	Botswana
Indonesian [ind]	Austronesian	Indonesia
Kogi [kog]	Chibchan	Colombia
Komnzo [tci]	Yam	Papua New Guinea
Matukar Panau [mjk]	Austronesian	Papua New Guinea
Vera'a [vra]	Austronesian	Vanuatu
Yurakaré [yuz]	Isolate	Bolivia

Table 1: Languages used in the present study.

Additionally, the task uses certain visual devices (speech and thought bubbles appear in 7 of the 16 depicted scenes), which may prompt task participants to discuss speech and thought more than they would in other contexts. For the present study, therefore, we added a "supplementary sub-corpus" for each of the 13 languages and annotated that data according to the same schema (§3.3), to check the representativeness of our SCOPIC findings. As far as practicable the supplementary sub-corpus is (roughly) equivalent in size to the corresponding SCOPIC data for each language.

Whereas the SCOPIC sub-corpora are parallax, and thus broadly equivalent, the non-SCOPIC sub-corpora differ for each language, according to the contingencies of what language-specific investigators have gathered or have access to. Among the genres represented (summarised by language in Table 2) are: Pear stories (represented in 7/13 of the languages), Frog Stories,[7] traditional stories (including folktales), autobiographical narratives, conversation, sociolinguistic interviews and TV Debates. Table 2 summarises the amounts of data in each sub-corpus and labels the kinds of data found in the non-SCOPIC sub-corpora.

[7] An anonymous reviewer correctly observes that the Pear Stories and Frog Stories are also parallax. However, there is nonetheless an important difference in the relation of the resulting monologues to the stimulus. In both Pear and Frog Stories, the story line is already given by the stimulus, namely the order of episodes. In the SCOPIC task speakers were free to construct a storyline according to their own logic (and after discussion amongst themselves), meaning that choices of the order of framing (and hence of givenness, for example) are freer.

Sub-corpus	Length	PF Annotations	Annotation equivalency
Time-based sub-corpus information			***n* per minute**
Arta SCOPIC	1:22:55	274	3.3
Arta Pear Story, Traditional stories, Autobiographical narratives	1:16:29	141	1.84
Avatime SCOPIC	1:55:51	147	1.27
Avatime Traditional stories, Pear Stories	40:24	138	3.42
Balinese SCOPIC	6:30:31	745	1.91
Balinese Traditional stories, Spontaneous dialogue	1:33:03	229	2.46
Dalabon SCOPIC	1:20:40	420	5.21
Dalabon Traditional story, Autobiographical narrative, Pear Story (commentary and recall)	1:17:53	306	3.93
English SCOPIC[8]	3:02:50	569	3.11
English Sydney Speaks (sociolinguistic interviews)[9]	2:14:29	462	3.44
G|ui SCOPIC	56:00	214	3.82
G|ui Pear story, Interview, Traditional Stories	55:13	399	7.23
Indonesian SCOPIC	1:05:13	249	3.82
Indonesian Pear Stories, Autobiographical narratives, Traditional stories	34:01	195	5.73
Komnzo SCOPIC	1:20:59	160	1.98
Komnzo Narrative, Conversational Narratives	1:12:21	219	3.03
Matukar Panau SCOPIC	2:12:08	449	3.4
Matukar Panau Frog Stories, Exposition, Autobiographical narratives	2:13:37	127	0.95
Vera'a SCOPIC	1:13:44	129	1.75
Vera'a Pear Stories, Traditional and modern narratives, Local history	41:28	66	1.59
Yurakaré SCOPIC	1:36:02	513	5.33

[8] Five of our English SCOPIC sessions were collected by Gabrielle Hodge and Kazuki Sekine as part of a bilingual, multimodal corpus for comparison with Auslan (Hodge et al. 2019). Our thanks to them for providing this data.

[9] Travis et al. 2023. Our thanks to Catherine Travis for providing access to this data.

Sub-corpus	Length	PF Annotations	Annotation equivalency
Time-based sub-corpus information			***n* per minute**
Yurakaré Traditional stories, Sociolinguistic and other interviews[10]	1:41:30	245	2.41
Word count based sub-corpus information			**Per 1,000 words**
German SCOPIC	14,567	572	39.27
German Teacher feedback, TV debate, Narrative, Interview from Datenbank für Gesprochenes Deutsch[11]	21,193	1,037	48.93
Kogi SCOPIC	6,111 words	177	28.96
Kogi Frog Story, Pear Story, Traditional Story, Autobiographical narrative	2,145 words	80	37.3

Table 2: Sub-corpora summary information (PF = Propositional framing).

Including this supplementary comparison corpus allows us to determine if there is the same amount of propositional framing in SCOPIC vs other sub-corpora, and if it is of a different kind. This determination allows us to assess both the validity and reliability of using SCOPIC to answer questions relating to the typology of social cognition.

While SCOPIC and non-SCOPIC corpora are of roughly the same order of magnitude, for various reasons they are not identical. In general, if there is a discrepancy, the SCOPIC corpus is bigger, reflecting the fact that we have been annotating that over many years. These differences in sub-corpus size will be smoothed out by normalisation, as outlined in §4.

3.3 Annotation schema

All material from both the SCOPIC and the supplementary corpora were transcribed and translated using ELAN, a software for annotating audiovisual data developed at

[10] Data collection funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project 275274422, reference number GI 1110/1-1) Global South Studies Center, University of Cologne, DobeS initiative of the Volkswagen Foundation (grant numbers 81821 and 83448). Also includes data sourced from van Gijn et al. (2011).

[11] Selections from the *Datenbank für Gesprochenes Deutsch* (DGD): *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK); *Deutsche Umgangssprachen: Pfeffer-Korpus* (PF); *Deutsche Mundarten: Zwirner-Korpus* (ZW), see IDS (n.d.).

The Language Archive, MPI Nijmegen, The Netherlands (e.g. Brugman & Russel 2004)[12] by the investigator(s) responsible within our team for that language, working with native language users if this was not their mother tongue. For each phenomenon of interest to our broader project we collectively developed an annotation schema, over a series of meetings of three continental "sub-teams" – Australian, Japanese and European – with Barth and Evans present at all of them and some other overlap of membership. This helped to support annotation reliability and typological categorical coverage across the sample of languages. Annotations cover a range of features relevant to social cognition, many irrelevant to the questions we are Investigating here, e.g. formulation of reference, use of kin terms; see Barth et al. (2021). For each of these features, values are annotated on a separate tier of our ELAN files. See Barth et al. (2024) and the appendix of Kimoto et al. (2024) for a full discussion of our annotation schema across all variables. Here we confine our discussion to the annotations relevant to complementation and the functions it expresses.

A fundamental question to resolve before we begin is whether complementation is defined structurally or functionally. Noonan (1985), in his classic typological overview of complementation, defines it as "the syntactic situation that arises when a notional sentence or predication is an argument of a predicate. For our purposes, a predication can be viewed as an argument of a predicate if it functions as the subject or object of that predicate" (Noonan 1985: 52).[13] But, after offering this syntactic definition, Noonan's chapter goes on to display some conceptual slippage between semantic and structural definitions. Thus on p. 64, picking up his earlier definition, he writes "in contrasting this *(universal) semantic characterization* [italics ours] with the surface characteristics of sentences containing complements…". Aikhenvald & Dixon (2006: 1), who maintain overtly that not all languages exhibit complementation structures, then propose the term *complementation strategies* for what happens in languages which, while lacking complementation structures, "still do have some grammatical mechanism for stating what a proposition is which is seen, heard, believed, known, liked, etc".

A fundamental assumption of typology is that we have to reckon with differential mappings of function onto form across languages. This creates a need to distinguish,

[12] Exceptions for supplementary corpora include Kogi and German data that were annotated based on textual transcriptions. Annotations were produced in either a text document or spreadsheet.

[13] Cf the definition in Aikhenvald & Dixon (2006: 4): "A complement clause has the following basic properties: (I) It has the internal constituent structure of a clause. (II) It functions as a core argument of a higher clause".

terminologically, between particular structures and the functions they express.[14] Now it often happens that there is terminological contamination between these two planes of linguistic analysis, as exemplified by the Noonan and Aikhenvald & Dixon quotes just given, but our preference is to keep the distinction as clear as possible by using entirely different terminology. We will therefore use the functional term *propositional framing* for the function and *complementation* for the structure as per Noonan's definition above.

The need for a comparable distinction between structure and function in the study of complementation has been repeatedly emphasised by scholars; in addition to the Noonan and Aikhenvald & Dixon passages above we can cite these lines from Deutscher's (2000) study of the evolution of complement structures in Akkadian:

> Different languages (or different diachronic stages of the same language) can use different structures to perform similar functions. For this reason, a valid comparison between languages needs to examine the role not only of complements, but also of other structures (such as parataxis), which can perform similar functions. This study differentiates between two concepts: 'complementation' (the embedding of a clause as an argument of a predicate), and the 'Functional Domain of Complementation', which includes complements as well as other strategies that perform similar functions (Deutscher 2000: 4).

Reflecting this need to distinguish claims about *structure* (here, complement structures) from those about *function* (e.g. expressing mental attitudes), our annotation scheme labels each relevant constructional token with a two-part annotation,[15] where the first part categorises the function and the second part its structure. Consider an annotation like UTT:COMP. This is to be understood as "utterance predicate [expressed by] complement structure"; it would be appropriate for an English sentence like *I told them that she arrived.* A complement structure could also be used for a probability judgment In English, e.g. *I guessed that she had arrived*

[14] Ideally, we would also have annotations for prosody, since there are languages like Teiwa (Sauerland et al. 2020) where complements appear to be paratactic if one confines oneself to morphosyntactic criteria, but can be argued to involve hypotactic structures once prosodic criteria are brought in. This would be an important extension of our investigations, but because comparisons across our set would only work if the prosodic analysis of each language in it was sufficiently advanced, we have been unable to integrate this into our annotations at this stage.

[15] This two-part schema, adopted for the present paper, is a simplified version of a more complex annotation schema used in other analyses by our team (Kimoto et al. 2024). In that schema, we also categorise the framing element (e.g. clause, particle), whether there is any connective present, and features of the content proposition (e.g. indirect speech). For current purposes we use a simplified version of this annotation, giving just the construction type and the functional domain.

or *It's likely that she has arrived*, and both would be annotated PROB:COMP. But an adverbial expression could also be used to express a comparable function, e.g. *She has probably arrived*, and this would then be annotated with the same function but with a different structure, as PROB:FUSE.

Our envelope of annotation possibilities, for propositional framing functions, takes in all functions which complementation structures express in at least some languages, as explored in Noonan's (1985) treatment. To assess the presence and distribution of propositional framing constructions across our data, we use the categories in Tables 3 and 4 respectively; for more detail see Kimoto et al. (2024) and Barth et al. (2024).

Abbrev	**Category**	**Example**
COM	Commentative	*I regretted that she had arrived.*
DES	Desiderative	*I hoped that she would arrive.*
FEAR	Fear	*I feared that she would arrive.*
IMM	Immediate perception	*I saw that she had arrived.*
KNO	Knowledge	*I knew that she had arrived.*
PRET	Pretence	*I pretended that she had arrived.*
PROB	Probability judgment	*I guessed that she had arrived.*
THINK	Thought	*I believed that she had arrived.*
UTT	Utterance	*I told them that she arrived.*

Table 3: Annotation categories for propositional framing functions.

Abbrev	**Structure**	**Example**
ADV	Adverbial clause (the framing element is in a subordinate clause to the clause expressing the content proposition)	*I am sorry, because the inspector has come.*
COMP	Complementation	*I regret that the inspector has come.*
COORD	Coordination	*I saw her and she was sleeping.*
FUSE	Proposition and frame fused into one clause: i.e., with a nominalised argument, verbal inflection or other bound morphemes, adverbs	*I regretted the inspector's arrival. / Maybe she is coming.*
INDP	Independent sentence with no framing element	*"She's come!"* [attributed speech uttered with no overt frame]
PARA	Parataxis	*I'm upset. The inspector has come.*
PRTH	Parenthetical	*The inspector, I can see, has arrived.*
SUB	Other subordination (including noun-modifying clauses)	*The picture of him arriving*

Table 4: Annotation categories for propositional framing structures.

Note that, to have a workable typology for comparison and to have sufficient tokens for statistical analysis, many of these categories are defined in a broad way. For example, FUSE includes bound adverbial prefixes (such as Dalabon *molkkunh-* in (7)) and other bound material such as desiderative inflections expressing ‘want to’, such as *-tai* in Japanese (*nometai* ‘wants to drink’). While unsatisfactory for a maximally delicate typology, such borderline cases, and the functions they express, are relatively rare in our corpus, so that these simplifications do not affect our overall findings.

4. Cross-corpus comparisons

4.1 Quantitative Results: correlation heatmap dendrogram

In this section we first investigate complementation and its functional equivalents through quantitative means, across both corpus types, then use more qualitative methods to probe the reasons for the few differences we find between the SCOPIC and other sub-corpora. Our quantitative analysis is a means of seeing how similarly different languages pattern based on the ways they pair such structures as complementation, parataxis, fused constructions and other means of framing propositions with the various functions identified above. Our analysis also assesses how and whether the data type (SCOPIC or other) changes how languages pattern.

Our data included 6,973 annotations across all corpora. We excluded 12 tokens due to low use of the COORD structures. Other structural categories have between 136 (PRTH) and 2,702 (COMP) tokens. After exclusions, there are 6,961 tokens.

Figure 2 shows the distribution of construction types for each language, by each sub-corpus. Our take-aways from this figure are that the SCOPIC data often increases the amount of propositional framing used (see the taller bars for Balinese, English, Matukar, *inter alia*), and that a more heterogenous distribution of construction types tends to be found in the SCOPIC sub-corpora (see German [DEU] non-SCOPIC having primarily complement clauses, but SCOPIC data having also many fused-inflectional constructions and subordinating constructions). These are signs of task validity: namely that the task encourages people to produce more structures relating to social cognition (including propositional framing) than we would otherwise be able to observe in recorded data.

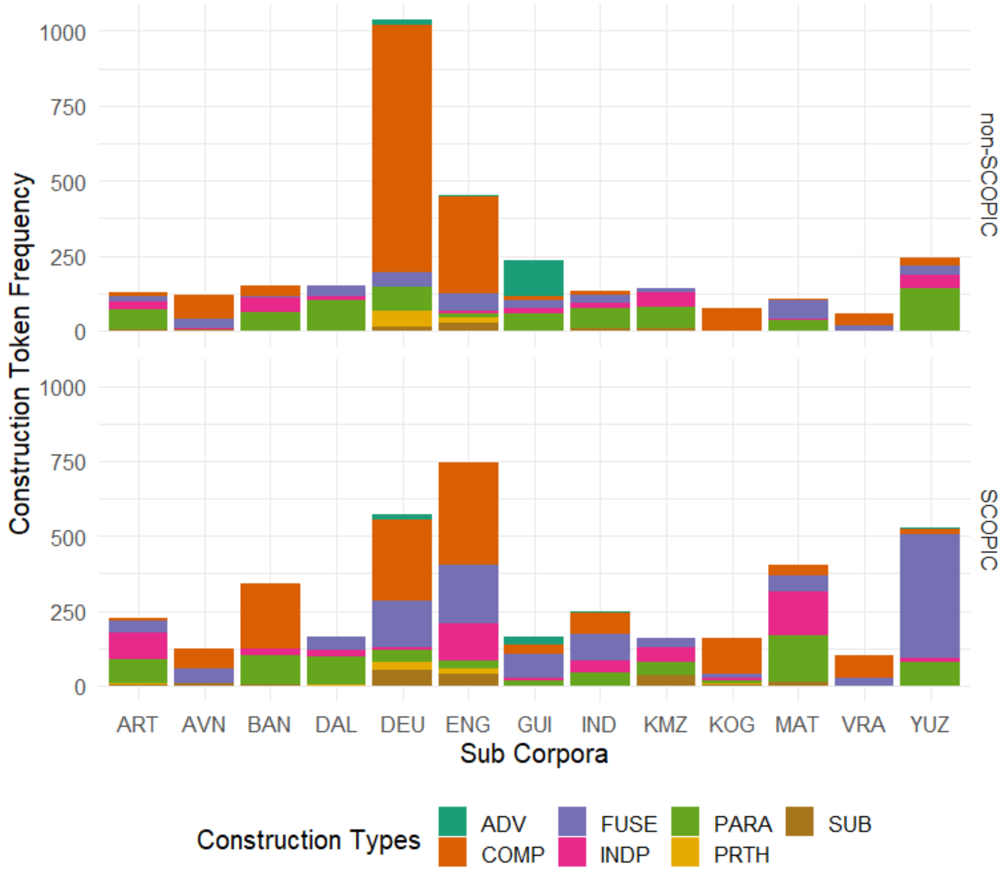

Figure 2: Construction types by language and sub-corpus.

Another important difference between the SCOPIC and non-SCOPIC data, particularly evident across quite a few languages (and most marked in Avatime, English, German, G|ui and Yurakaré) is the higher incidence of the FUSE structure in the SCOPIC sub-corpora (as shown by the increase in purple slices of the bar graphs). This reflects one important way in which the SCOPIC task does skew the data: during the SCOPIC task people use expressions of uncertainty that are integrated into the clause, like the adverbial expressions ENG *maybe,* DEU *vielleicht* 'possibly, maybe' when people are hazarding guesses about what particular cards depict and how they relate to each other (e.g. whether two cards contain the same character). This is reflected in Figure 3 which shows considerably more PROB tokens (light orange) for the SCOPIC data across many of the languages. However, we see similarities across the SCOPIC/non-SCOPIC data in that there often are a large amount of reported utterance tokens (dark orange THINK and purple UTT) and a small amount of pretence (red) and knowledge (pink) tokens.

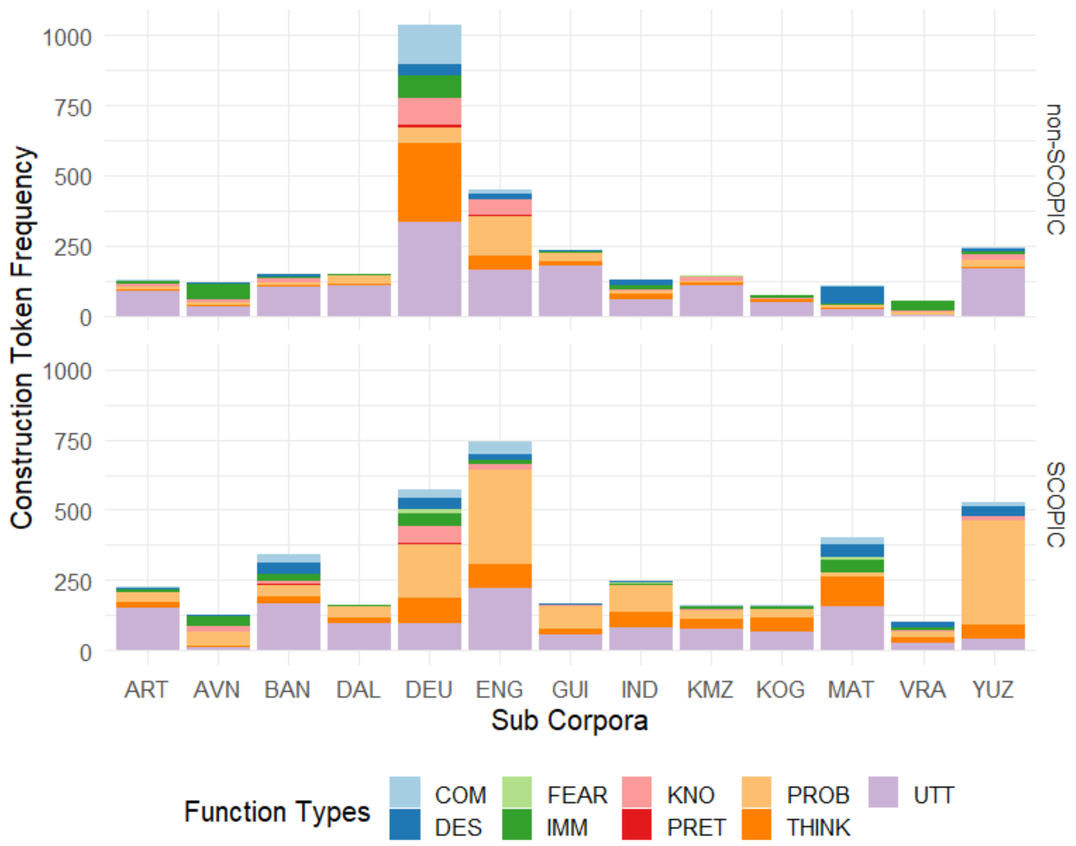

Figure 3: Function types by language and sub-corpus.

Finally in our comparison of the bar charts, we notice that for some languages, there is a clear mapping between construction type and function type. Take Matukar, for instance, which has very similar type distributions in the construction and function types as UTT and THINK functions map directly to PARA and INDP constructions, DES functions map directly to FUSE constructions, and COM functions map directly to COMP constructions. This is not the case for our COMP construction heavy languages like German, English and Kogi, where the orange COMP slices of the bars in Figure 2 are very dominant, even when there is a more heterogenous distribution of function types. This is a sign that complementation is being used for a wide range of functions in these languages.

We now turn to the question of how similar and dissimilar the languages and sub-corpora are from each other based on their distribution of construction types. Investigating this question will help show whether complementation is predominant or not, and if not, what constructional alternatives there are available. Clustering analysis gives us insight into what kinds of typological usage profiles we find in the different sub-corpora so we can consider its causes and the impact of the evolution of language complexity in terms of propositional framing.

We structure our data through "profiles of use". This means that for each sub-corpus we calculate a proportion of construction types used. We then evaluate how different each sub-corpus' profile of use is through a clustering analysis and heat map dendrogram created with the *R* package (R core Team 2023) *pheatmap* (Kolde 2019). We find four clear clusters in our data based on the proportion of use of each construction type. Figure 4 shows a heat map dendrogram with annotations labelling the clusters.

Heat map dendrograms are useful visualisations of differences as they combine two kinds of analyses: correlation heat maps and clustering dendrograms. The clustering dendrogram is a tree-structured representation of the data and how it clusters. Items (here sub-corpora) in each cluster are more similar to each other than to items from other clusters. This clustering is represented by the branches on the side and top of the heat map in Figure 4. Each branch shows which construction types (top dendrogram) and sub-corpora (side dendrogram) are best grouped together. Horizontal and vertical cuts of the data show where the strongest branching is. Lines connect the nodes that form clusters (King 2015). It is the clustering analysis that determines the order of constructions and sub-corpora along the x-axis and y-axis of these figures. The package *NbClust* (Charrad et al. 2014) was used to determine that the ideal number of clusters for the heatmap was four. Our analysis of this clustering leads us to characterise the clusters as [1] – COMP cluster, [2] - PARA/INDP cluster, [3] – FUSE/ADV cluster.

The correlation heat map uses colour to show how strongly entities in a matrix are associated. Each cell in our matrix represents how strongly a construction type and a sub-corpus are associated. The number in each cell reflects this association size. In our heat map in Figure 4, red is for "hot" to show a strong positive association between the sub-corpus and the construction type (positive value), and blue is for "cold" to show a strong negative association (negative value). Yellow shows no substantial association. As shown by labels on the x-axis in Figure 4, we are essentially counting the number of times each construction type was used within a sub-corpus to frame propositions. We then normalise these counts by centring and scaling the values (see Lucas et al. 2020). We group these measures by sub-corpus listed on the y-axis.

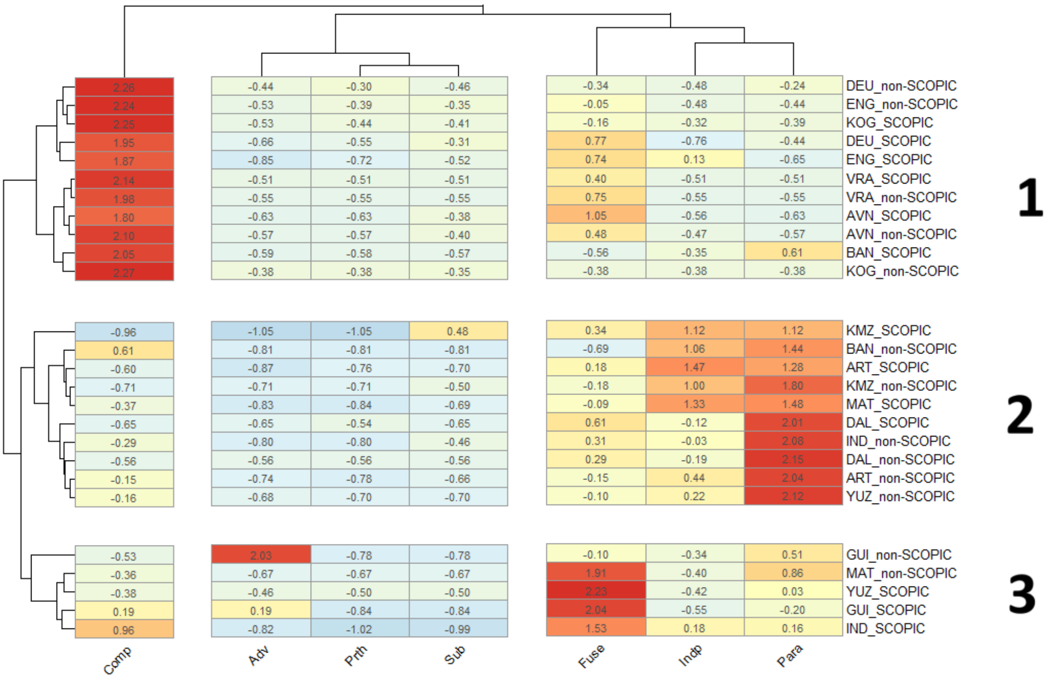

Figure 4: Heatmap dendrogram showing clusters of sub-corpora by the association with each construction type.

We now turn to the examination of each of the three clusters. More than half of the SCOPIC and non-SCOPIC sub-corpora for each language fall within the same cluster; we call these *sub-corpora pairs*.

Cluster 1 (COMP cluster) has sub-corpora pairs from German (DEU), English (ENG), Kogi (KOG), Vera'a (VRA) and Avatime (AVN), reflecting that these languages use high rates of complementation in all their data types. Balinese SCOPIC data is also in the COMP cluster. As opposed to the Balinese non-SCOPIC data, there are tokens from a wider range of functions present. These functions (DES, IMM, PROB) all map to COMP constructions, as well as about half of the reported utterances. This clustering reflects a pattern seen in the descriptive bar charts: if a language uses complementation structures, that structure is used for a wide variety of functions.

Cluster 2 (PARA/INDP cluster) has sub-corpora pairs for Dalabon (DAL), Komnzo (KMZ) and Arta (ART), reflecting a high amount of parataxis in all their data types. The cluster also has non-SCOPIC data from Yurakaré (YUZ), Indonesian (IND) and Balinese (BAN) and SCOPIC data from Matukar (MAT). For these four languages, there is a higher proportion of UTT and THINK functions than any other functions in their cluster 1 sub-corpora. This maps onto higher amounts of INDP and PARA constructions. Therefore, it is the function of reported utterance (expressed paratactically in these languages) and its high presence in the sub-corpora that drives the clustering for the latter four languages.

Cluster 3 (ADV/FUSE cluster) shows three SCOPIC sub-corpora from Yurakaré and Indonesian. These sub-corpora showed a higher amount of PROB function tokens, which mostly mapped to the FUSE construction type by using adverbs to modify/frame propositional content. The Matukar (discussed further in section 4.2.1) non-SCOPIC sub-corpus is clustered in the FUSE group due to a higher proportion of DESID functions in that corpus, which map onto the FUSE construction type through the irrealis, desiderative inflectional suffix (i.e. proposition and frame within a single clause). There is a sub-corpus pair from G|ui in cluster 3, reflecting that neither COMP nor PARA structures are prevalent in the language. The G|ui SCOPIC sub-corpus, similar to Yurakaré and Indonesian, has a high amount of FUSE constructions from the PROB functions. The G|ui non-SCOPIC sub-corpus is the only sub-corpus with many ADV structures, these are due to that structure being mapped to THINK/UTT functions from a high expression of reported utterance in the G|ui traditional stories.

4.2 Qualitative Results: sub-corpora differences and similarities

As mentioned above, there are some cases where there are differences in construction types between the sub-corpora, within a single language. We have found through qualitative analysis that this reflects a difference in the functions expressed in the sub-corpora. Namely, some constructions do not regularly occur without expressing a particular meaning, and the difference emerges in our quantitative counts, as shown in Figure 4. We follow this below with a discussion of distributions in three languages: English (both sub- corpora in cluster 2 with high rates of complementation), Dalabon (both sub-corpora in cluster 1 with low rates of complementation) and Matukar Panau (data split across clusters 1 and 3).

4.2.1 Sub-corpora differences: Matukar Panau

Matukar Panau has few complement constructions, parentheticals, adverbials or other subordination strategies for propositional framing. While Matukar Panau in general has considerable subordination, at least on the not uncontroversial assumption that clause chaining and serial verb constructions are types of subordination, these are not used for layering ideas. Rather, the subordination is used to show temporal links between actions (clause chaining like 'go, see, then stay'), or composite events (serial verb constructions like 'meet-eat' meaning 'feast'). Propositional framing subordination is fairly rare no matter the sub-corpora of Matukar Panau. To frame or comment on ideas, Matukar Panau uses paratactic and independent constructions, especially in the SCOPIC data. In the non-SCOPIC data, we see considerable inflectional construction usage in the frog stories, exposition and autobiographical narratives. We find that this is directly related to the semantic types used in the sub-corpora.

A common functional type across the sub-corpora is the desiderative. The sole FUSE propositional framing usage is for desideratives through an inflectional verbal suffix (as in Example 8). Additionally, in both sub-corpora, quoted speech and thought are expressed through independent and paratactic constructions. The reason the SCOPIC data is in a different cluster in our clustering analysis is that there is simply a lot more quotation in the SCOPIC texts (as in Example 9), reflecting their different profile of use. Therefore, the difference is not in the kind of syntax, but in the frequency of its usage. The distributions of the main kinds of constructions and their associated functions are laid out in Table 5. Other construction types are too infrequent to be worthwhile including here. Matukar Panau does have a (quasi-)complement clause structure, but it is infrequent and is restricted to commentative functions with fully inflected verbal clauses framed by non-verbal predication as in (10).

Data	Desiderative	Quoted thought	Quoted speech
SCOPIC Para		51	87
SCOPIC Indp		73	79
SCOPIC Fused	41	1	
Frog Story Para			1
Frog Story Indp			
Frog Story Fused	6		
Autobio Narrative Para		3	26
Autobio Narrative Indp		7	8
Autobio Narrative Fused	15		
Exposition Para			1
Exposition Indp			
Exposition Fused	39		

Table 5: Primary Matukar Panau Construction-Function distribution by text type.

(8) Matukar Panau (Tomas Taleo Kreno – DGB1-2013_031-Tomas_Taleu_Kreno_Frog_Story – 1:02.5-1:09)

gaun	*bab-bai*	*bul-do*	*aim*	*main*	*tuli-nge*
dog	bark-I.IRR.DESID	try-R.D	boy	TOP	say-I.R.PFV

“awa-m	*tau!”*
mouth-2SG	shut

‘The dog wanted to start barking and the boy said “shut up!”’

(9) Matukar Panau (Tomas Taleo Kreno – SocCog-mjk01-tk_jb_1 – 5:18.6-5:23.6)

garma-n	*alo=a*	*suse-nge*	*so-nge*
head-3SG	back=LOC	thread-D.SEQ	come-R.D

so-nge	*tamat*	*numa-n=te*
come-D.SEQ	man	hand-3SG=LOC

“so-ndop	*ngau=da*	*t-a”*	*i*	*bal-e*
come-D.IRR	1SG=COM	1PL.EXCL.SBJ-go	3	say-I.R.PFV

‘He tied his hair back and came, the man took her hand and he said “come, let’s go (lit: with me, we go)!”’

(10) Matukar Panau (Mingkui Agid – SocCog-mjk10-ckd_ma_2 – 4:39.1-4:40.6)

hum-e	*main*	*uyan*	*ti*
hit.PL-I.R.PFV	COMP	good	NEG

‘His hitting them is not good’

4.2.2 Sub-corpora similarities: English

English (Table 6) has a good number of complement clauses, and they are used consistently across a range of functions. The comparative corpus is the Sydney Speaks corpus (Travis et al. 2023) and consists of sociological interviews. The most frequent functions in both text types were knowledge, probability, quoted thought and quoted speech. We also see the same pattern of construction types across the semantic functions in both text types: complementation for all functions, fused for probability and some independent (unframed) quoted speech. Other combinations of constructions and functions are infrequent and not different enough to cause the text types to appear in different clusters in our clustering analysis.

Data	Knowledge	Probability	Quoted thought	Quoted speech
SCOPIC Indp		2	1	57
SCOPIC Fused		166	1	10
SCOPIC Complement	14	110	35	140
Sydney Speaks Indp				9
Sydney Speaks Fused		56		
Sydney Speaks Complement	26	76	47	53

Table 6: Primary English Construction-Function distribution by text type.

Where we do see some differences is in the specific form of complementation for quoted speech. In our SCOPIC data, the framed quote is often indirect rather than direct. In the Sydney Speaks data, it is more likely to be direct. Further, in the Sydney Speaks data, the connective between clauses is more often *like* than *that.* Additionally, in our SCOPIC data the quotee is always a specific and identifiable referent (example (11)). In the Sydney Speaks data, we observe that the quotee in the framing clause is sometimes a non-specific/impersonal referent as in ***it** was like* or ***stuff** was like* (as in example (12)). This is an indication that complementation/quotation in this construction is still an area that is undergoing grammaticalization, since the Sydney Speaks interviewees in our data are younger than the SCOPIC participants[16].

(11) English (Kat – SocCog-eng06-BK_English_Story_Oct3118pm – 24:25-24:29.7)
And he's like "Julie you've been fooling in with the shop assistant"

(12) English (SydS_AYF – SydS_AYF_128 – 9:48.5-9:53.2)
*it can't just go straight to her because things are piling up and **stuff was like "wa!"***

[16] Note that all of our English speaking participants are from Australia and speak an Australian variety of English. Therefore this more likely to be a generational change than a dialect difference.

4.2.3 Sub-corpora similarities: Dalabon

When we compare the texts making up the non-SCOPIC Dalabon sub-corpus we see a very clear pattern: by far the most common propositional framing use in any sub-corpus is paratactic quoted speech (13). Additionally, we see some fused constructions for probability in almost all of the texts, in the form of the particle *kardû* 'maybe' (14, 16, both from the Pear Story). Quoted thought is not common outside the SCOPIC sub-corpus, though the boundaries between quoted speech and quoted thought-as-speech can be hard to draw (15, 16)[17] but in the SCOPIC data and the Pear Story data from the non-SCOPIC corpus it follows similar distributions as quoted speech. Figures illustrating the comparison between the SCOPIC and non-SCOPIC data for Dalabon are given in Table 7.

(13) Dalabon (MP Pear Story Recall 0.2.43-0.2.48.5)

ka-h-na-ng *"Ngale!* *kardû*
3SG.SBJ > 3SG.OBJ-R-see-PST.PFV hey maybe

nga-h-dja-lng-karrû-bakm-inj"
1SG.SBJ-R-just-SEQ-leg-break- PST.PFV

'He looked (and said/thought/saw): "Hey, maybe I've broken my leg just now."' (Alt. transl.: 'He saw that maybe he had broken his leg.')

(14) Dalabon (MP Pear Story Commentary 01.31-01.33)

kardû *ba-ru-rr-inj*
maybe 3SG.SBJ.PST-rub-RR-PST.PFV
'Maybe he rubbed himself.' (i.e. rubbed his leg because he fell).

(15) Dalabon (MP Pear Story Commentary 01.33-01.35)

"ngaleh! *nga-h...* *nga-h-dordm-iyan!"* *ka-h-yin-inj*
hey 1SG.SBJ-R 1SG.SBJ-R-go.up-FUT 3SG.SBJ-R-say/do-PST.PFV
'"Hey, I'll climb up" he said/went/thought.'

[17] Cf Reesink (1993) who discusses this issue in Papuan languages.

(16) Dalabon (MP Pear Story Commentary 3.57-04.04)

kah...	*kardû,*	*wonarr-inj*	*“ngale*	*kardu*	*nga-h-lng-darrû...*
3SG.SBJ	maybe	think-PST.PFV	hey	maybe	1SG.SBJ-R-SEQ-leg

darru-bakm-inj	*bah*	*kahke*	*ka-h-dja-mon*
leg-break-PST.PFV	but	nothing	3SG.SBJ-R-just-good

‘Maybe he thinks “hey! maybe I’ve broken my... my leg, but nothing, it’s OK”’.

There are interconnected issues of lexicography, translation and corpus analysis here, well illustrated by (16). The stem *wonarr-* ‘think’ is formally the reflexive form of *wona* ‘hear’, so ‘to hear oneself’, a common pattern in Australian languages (Evans & Wilkins 2000: 571, which contains another and contextually comparable use of the same verb). So, one more literal translation of (16) would be ‘Maybe he heard himself saying: “Hey! Maybe I’ve broken my leg”’ while a less literal one, taking the semantic transition to ‘think’ into account, is ‘Maybe he thinks “Hey! Maybe I’ve broken my leg.”’ Examples like this illustrate the difficulty of reaching an analytic decision about whether to treat them as quoted speech or quoted thought, but at the same time show how natural it is to adopt paratactic quoted-speech constructions for representing mental states.

Data	Probability	Quoted thought	Quoted speech
SCOPIC Para		47	272
SCOPIC Indp		5	34
SCOPIC Fused	44	6	
Pear Story Para		3	22
Pear Story Indp			11
Pear Story Fused	1		
Autobio Narrative Para		1	63
Autobio Narrative Indp			6
Autobio Narrative Fused	20	2	
Event Recall Para			5
Event Recall Indp			
Event Recall Fused			
Traditional Story Para			138
Traditional Story Indp			3
Traditional Story Fused	17	1	

Table 7: Primary Dalabon Construction-Function distribution by text type.

5. Conclusions and implications

To our opening question, regarding whether complementation constructions are universal, our study suggests a negative answer. They are certainly common, occurring as the dominant structural type for both corpora in close to half the languages of our sample – Avatime, English, German, Kogi and Vera'a – and as the dominant type in the non-SCOPIC corpus in one more (Balinese). And they are attested, at least once in both corpora, in all languages except Dalabon. On the other hand, they are entirely absent for one language in our corpus (Dalabon; both corpus types[18]), entirely absent from the non-SCOPIC corpus in Komnzo, and only occur at low frequencies in both corpora for Arta, G|ui, Indonesian, Matukar Panau and Yurakaré, and low frequencies in the non-SCOPIC corpus in Balinese. Paratactic structures dominate in Dalabon and Arta, as well as the non-SCOPIC corpora for Indonesian, Komnzo and Yurakaré. For G|ui, other subordination strategies predominate, and fused strategies are dominant in for several languages in the SCOPIC corpus (English, German, Indonesian, Yurakaré). As discussed in §4, however, the elevated occurrence of fused strategies in these latter languages reflects the large number of words like *maybe* elicited by people qualifying their speculative interpretations of pictures in the task.

Our finding regarding the non-universality of complementation fits in with a number of claims by other scholars. For the oldest varieties of Akkadian, spoken around 4,500 years ago, Deutscher (2000) argues that complementation was absent and that it was only later that the erstwhile causal subordinator *kīma* gradually developed into a complementiser, via a reanalysis path from causal adverbial clause structures of the type *He said/spoke to the governor* ***because*** *(k-marker) the barley was not collected* to complement structure of the type *He said/spoke to the governor* ***that*** *(k-marker) the barley was not collected.* Givón (1991) proposes a somewhat similar

[18] Our assertion about the lack of complementation in Dalabon is based on the two corpora in our study. A more far-reaching survey of Dalabon grammar turns up one highly specialised construction that could be analysed as complementation: 'want' verbs with different or partially disjoint subjects, which are clearly conventionalised biclausal constructions where the first 'want' verb is marked with a benefactive applicative which uses indirect object marking to index the subject of the complement. See Evans (2006, 2021). These constructions are extremely marginal, as indicated by their complete absence from the two corpora reported on here and their virtual confinement to elicited settings. In another Dalabon corpus of around 60 hours (Ponsonnet 2013), Maia Ponsonnet (emails to NE, 7/1/2025 and 10/1/2025) found just three examples.

scenario for the Biblical Hebrew complementiser *kī*, but this time revolving around framing predicates like *be happy / regret*; again a reanalysis from *be happy because X* to *be happy that X* is a small step semantically, providing a clear bridging context for the structural reanalysis from adverbial clause to complement clause.

More recently Hernáiz Gomez (2024), drawing on a larger corpus of the earliest forms of Akkadian, disputes Deutscher's claim that complement structures were entirely absent, but nonetheless goes on to show that in a number of Semitic languages with lengthy written traditions what are now clear complement structures originated as similative manner expressions.

These arguments gel with diachronic studies from elsewhere in the world that show how complement structures can emerge, by such means as the grammaticalisation of report verbs into quote markers and complementisers in the Austronesian languages Tukang Besi and Buru (Klamer 2000), the progression from paratactic to hypotactic structures (Harris & Campbell 1995), the tightening up of intonational links from *unbound* to *bound* (Diessel & Hetterle 2006) and the reanalysis of clausal or intonational boundaries so that *'He said this/that. "X"'* becomes *'He said [that X]'*, as shown for Mohawk by Mithun (2025).

Regarding recursive structures more generally, Widmer et al. (2017: 799), in their careful diachronic study of Indo-European, have shown, for recursive NP embedding, that "every type of NP embedding – genitives, adjectivisers, adpositions, head marking, or juxtaposition – is unavailable for syntactic recursion in at least one attested language. In addition, attested pathways of change show that NP types that allow recursion can emerge and disappear in less than 1,000 years". The net effect of these studies is to show that complementation, like other types of recursion, is by no means a universal structure: rather, it is something that evolves and sometimes disappears over time.

The reader will have noticed that, in our discussion of how complement structures emerge from various other structural sources (causal adverbs in Ancient Akkadian, speech reports in the cases examined by Klamer and Mithun), the material corresponding to the complement in a language like English is treated as reported quotation. Revealingly, this is why Dalabon, the most complement-averse language in our sample, makes such extensive use of parataxis: direct speech, whether actual or represented/projected, is combined with a wide range of framing verbs to convey the equivalent of complement structures in a language like English; in our corpus these include such verbs as *bengdinj* ‘was thinking’, *yolhwehmun* ‘feels bad, worries’, *bengkang* ‘thought’, *kurnh-bengkabengkang* ‘thought, worried’, as well as other nominal framing devices like *men-no* ‘his/her mind’ (see further discussion and examples in Rumsey et al. 2022). This suggests that the most primal origins for such structures are to be found, not in syntax, but in potentially recursive embeddings of passages of quotes within one another, in other words intertextual embedding at the narrative or discourse rather than the syntactic level. This is the argument advanced in an important recent article by Spronck & Casartelli (2021: 19):

> The type of linguistic structures specifically dedicated to this task are reported speech. If linguistic reflexivity, that is, thinking and talking about language, is at the heart of the complexification of grammar, reported speech is at the heart of language evolution, which would at once explain its universality in the languages of the world and its relation to grammatical categories.

The present study, by pinpointing the many constructional alternatives to complementation that exist across a parallax corpus, helps clarify why it is not a necessary structure, since languages can employ many other means to realise the same communicative goal.

Grzech & Bergqvist (2025), in their introduction to a volume on the typology of evidentials and epistemics, highlight the growing realisation within linguistics that to truly understand language, communication and cognition, we must look “beyond single minds toward cognition as a process involving interacting minds” (Dingemanse et al. 2023: 1), and our need for methods that allow us to do this in a systematic cross-linguistic mode. They particularly mention the importance of Corpus Based Typology (Schnell & Schiborr 2022, Levshina 2022) as a way of capturing “intra-linguistic variation in language use and its relation to aspects of language systems” (Schnell et al. 2021: 6), and the need to ensure that such corpora, of which SCOPIC is an example, “are interactive and purposefully designed to explore social cognition, and allow an insight into how knowledge rights and obligations are negotiated in dialogic interaction” (Grzech & Bergqvist 2025: 15). In this article we have argued that appropriately designed corpora can indeed furnish information relevant to social cognition and can contain sufficiently high occurrence rates in domains of interest to reach statistically robust conclusions. At the same time, these results are compatible with the patterning found in less targeted corpora.

More specifically, using such corpora to examine the cross-linguistic occurrence of complementation or its functional equivalents allows us to see that complementation structures, though common, are not universal and that there exist a significant number of structural alternatives to them, most importantly paratactic constructions involving represented speech (Kimoto et al. 2024). These paratactic constructions are employed not just for speech per se, but for "internal mono/dialogue" accompanying thoughts, memories, intentions and perceptions. By showing that it is these structures, rather than syntactically specialised complementation constructions, which are truly ubiquitous, we are drawn back to the Bakhtinian insight that it is *raznorečie* or heteroglossia, the threading together of different people's words, which is what is truly universal in how we construct infinitely large and complex linguistic units from finite numbers of words and construction types. At the same time, the common grammaticalisation pathway by which complementisers can evolve from elements introducing quoted passages, whether verbs (e.g. Klamer 2000 on Austronesian), demonstratives (Mithun 2025 on Mohawk), similative/manner expressions (Hernáiz Gomez 2024 on Semitic) or causal subordinators (Deutscher 2000 on Akkadian), indicates how it is possible for the widespread occurrence of complementation constructions across languages to be linked to its roots in quoted speech.

Acknowledgements

This work was funded by the Australian Research Council (ARC; DP0878126, DP130101655, DP140102124, FL130100111), the *Anneliese-Maier Forschungspreis*, awarded to Evans by the Alexander von Humboldt Foundation and the German Federal Ministry of Education and Research, the *ARC Centre of Excellence for the Dynamics of Language (CoEDL)* (ARC; CE140100041), the *Deutsche Forschungsgemeinschaft* (DFG; 275274422, 417675039), the *Japan Society for the Promotion of Science* (Kakenhi grants JP18K00582 and JP22K00536), the *Marcus and Amalia Wallenbergs Minnesfond* (MAW; 2017.0081), the *Swedish Research Council* (VR; 2017-01969 and VR; 2020-01581), the *Swiss National Science Foundation* (P0BEP1_165335), the *UK Arts and Humanities Research Council* (AHRC; AH/N00924X/1), the Global South Studies Center, University of Cologne, and the Volkswagen Foundation DoBeS program (VW: 81821 83448, 85606, 86357). We thank the above-named institutions for their generous support of our research. We also thank Silvia Ballarè, Simone Mattiola, Nicola Grandi & Caterina Mauri, convenors of the conference at which these ideas were presented, *Naturally occurring data in and beyond linguistic typology* (Bologna, 18-19 May 2023), for their kind invitation to present there, two anonymous referees for their useful critical comments on an earlier version of this paper, Keira Mullan for assistance with the formatting, and Rodrigo Hernaiz Gomez and Marianne Mithun for further useful discussion and references. Above all, we thank all the language speakers who participated in this research for their friendship and help.

Abbreviations

1 = first person
2 = second person
3 = third person
A = agent
ACC = accusative
BEN = benefactive
COMP = complementiser
COP = copula
D = dependent
DEM = demonstrative
DESID = desiderative
DIS = disharmonic
DU = dual
EXCL = exclusive
FP = final particle
FUT = future
I = independent
IPFV = imperfective
IRR = irrealis
LOC = locative
NEG = negation
NOM = nominative
OBJ = object
PFV = perfective
PL = plural
POL = politeness marker
POSSD = possessed noun
PROG = progressive
PST = past
QUOT = quotative
R = realis
RR = reflexive/reciprocal
SBJ = subject
SEQ = sequential
SG = singular
TOP = topic

References

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2006. Introduction. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Complementation: a cross-linguistic typology*, 1–48. Oxford: Oxford University Press.

Barth, Danielle & Nicholas Evans (eds.). 2017a. The Social Cognition Parallax Corpus (SCOPIC). *Language Documentation and Conservation Special Publication* 12.

Barth, Danielle & Nicholas Evans. 2017b. The social cognition parallax corpus (SCOPIC): design and overview. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 1–21.

Barth, Danielle, Nicholas Evans, I Wayan Arka, Henrik Bergqvist, Diana Forker, Sonja Gipper, Gabrielle Hodge, Eri Kashima, Yuki Kasuga, Carine Kawakami, Yukinori Kimoto, Dominique Knuchel, Norikazu Kogura, Keita Kurabe, John Mansfield, Heiko Narrog, Desak P. Eka Pratiwi, Saskia van Putten, Chikako Senge & Olena Tykhostup. 2021. Language vs. individuals in cross-linguistic corpus typology. In Stefan Schnell, Geoffrey Haig & Frank Seifart (eds.), *Doing corpus-based typology with spoken language corpora: State of the art* (Language Documentation & Conservation Special Publication 25). 1–56.

Barth, Danielle, Nicholas Evans, Sonja Gipper, Stefan Schnell, Henrik Bergqvist, Menguistu Amberber, I Wayan Arka, Christian Döhler, Diana Forker, Volker Gast, Dolgor Guntsetseg, Gabrielle Hodge, Eri Kashima, Yukinori Kimoto, Norikazu Kogura, Dominique Knuchel, Inge Kral, Keita Kurabe, John Mansfield, Heiko Narrog, Desak Putu Eka Pratiwi, Hiroki Nomoto, Seongha Rhee, Alan Rumsey, Lila San Roque, Andrea C. Schalley, Asako Shiohara, Elena Skribnik, Olena Tykhostup, Saskia van Putten & Yanti. 2024. The Social Cognition Parallax Interview Corpus (SCOPIC) Project Guidelines. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 163–237.

Brugman, Hennie & Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2065–2068. Lisbon: European Language Resources Association (ELRA).

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau & Azam Niknafs. 2014. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6). 1–36.

Chomsky, Noam. 1957. *Syntactic structures.* The Hague: Mouton.

Deutscher, Guy. 2000. *Syntactic change in Akkadian.* Oxford: Oxford University Press.

De Villiers, Jill. 2000. Language and theory of mind: what are the developmental relationships?. In Simon Baron-Cohen, Helen Tager-Flusberg & Donald J. Cohen (eds.), *Understanding other minds: perspectives from developmental cognitive neuroscience* 2nd edn., 83–123. New York: Oxford University Press.

De Villiers, Jill G. & Peter A. De Villiers. 2003. Language for Thought: Coming to understand false beliefs. In Dedre Gentner & Susan Goldin-Meadow (eds.), *Language in Mind*. 335-384. Cambridge: MIT Press.

De Villiers, Jill G. & Peter A. De Villiers. 2014. The role of language in Theory of Mind Development. *Top Lang Disorders* 34(4). 313–328.

De Villiers, Jill G. & Jennie E. Pyers. 2002. Complements to cognition: the relationship between complex syntax and false-belief understanding. *Cognitive Development* 17(1). 1037–1060.

Diessel, Holger & Katja Hetterle. 2006. Causal clauses: a cross-linguistic investigation of their structure, meaning and use. In Peter Siemund (ed.), *Linguistic Universals and Language Variation*, 21–52. Berlin: Mouton de Gruyter.

Dingemanse, Mark, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K. Ameka, Abeba Birhane, Dimitris Bolis, Justine Cassell, Rebecca Clift, Elena Cuffari, Hanne De Jaegher, Catarina Dutilh Novaes, N. J. Enfield, Riccardo Fusaroli, Eleni Gregoromichelaki, Edwin Hutchins, Ivana Konvalinka, Damian Milton, Joanna Rączaszek-Leonardi, Vasudevi Reddy, Federico Rossano, David Schlangen, Joanna Seibt, Elizabeth Stokoe, Lucy Suchman, Cordula Vesper, Thalia Wheatley, Martina Wiltschko. 2023. Beyond Single-Mindedness: A Figure-Ground Reversal for the Cognitive Sciences. *Cognitive Science* 47(1). e13230.

Evans, Nicholas. 2006. Who said polysynthetic languages avoid subordination? Multiple subordination strategies in Dalabon. *Australian Journal of Linguistics* 26(1). 31–58.

Evans, Nicholas. 2021. Social cognition in Dalabon. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 22–84.

Evans, Nicholas, Francesca Merlan & Maggie Tukumba. 2004. *A first dictionary of Dalabon (Ngalkbon)*. Winnellie: Bawinanga Aboriginal Corporation.

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

Frajzyngier, Zygmunt. 1984. On the Origin of say and se as complementizers in Black-English and English-based creoles. *American Speech* 59(3). 207–210.

Givón, Thomas. 1991. The evolution of dependent clause morpho-syntax in Biblical Hebrew. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to Grammaticalization: Volume II. Types of grammatical markers*, 257–310. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Grzech, Karolina. & Henrik Bergqvist, 2025. Epistemicity in language: current horizons, future directions. In Karolina Grzech & Henrik Bergqvist (eds.), *Expanding the Boundaries of Epistemicity: Epistemic Modality, Evidentiality, and Beyond*, 1–30. Berlin: De Gruyter Mouton.

Harris, Alice & Lyle Campbell. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Hernáiz Gomez, Rodrigo. 2024. The grammaticalization of manner expressions into complementizers: insights from Semitic languages. *Linguistics: An Interdisciplinary Journal of the Language Sciences* 62(3). 617–651.

Hodge, Gabrielle, Kazuki Sekine, Adam Schembri & Trevor Johnston. 2019. Comparing signers and speakers: Building a directly comparable corpus of Auslan and Australian English. *Corpora* 14(1). 63–76.

Kimoto, Yukinori, Asako Shiohara, Danielle Barth, Nicholas Evans, Norikazu Kogura, I Wayan Arka, Desak Putu Eka Pratiwi, Yuki Kasuga, Carine Kawakami, Keita Kurabe, Heiko Narrog, Hiroki Nomoto, Hitomi Ono, Alan Rumsey, Andrea C. Schalley, Yanti, Akiko Yokoyama. 2024. Syntactic embedding or parataxis? Corpus-based typology of complementation in language use. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 126–162.

King, Ronald S. 2015. *Cluster analysis and data mining: An introduction.* Dulles: Mercury Learning and Information.

Klamer, Marian. 2000. How report verbs become quote markers and complementisers. *Lingua* 110(2). 69–98.

Kolde, Raivo. 2019. pheatmap: Pretty heatmaps (R package version 1.0.12). Available online at: https://CRAN.R-project.org/package=pheatmap (Accessed 2025.12.28).

Levshina, Natalia. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology* 26(1). 129–160.

Lucas, Carolina, Patrick Wong, Jon Klein, Tiago B. R. Castro, Julio Silva, Maria Sundaram, Mallory K. Ellingson, Tianyang Mao, Ji Eun Oh, Benjamin Israelow, Takehiro Takahashi, Maria Tokuyama, Peiwen Lu, Arvind Venkataraman, Annsea Park, Subhasis Mohanty, Haowei Wang, Anne L.Wyllie, Chantal B. F. Vogels, Rebecca Earnest, Sarah Lapidus, Isabel M. Ott, Adam J. Moore, M. Catherine Muenker, John B. Fournier, Melissa Campbell, Camila D. Odio, Arnau Casanovas-Massana, Yale IMPACT Team, Roy Herbst, Albert C. Shaw, Ruslan Medzhitov, Wade L. Schulz, Nathan D. Grubaugh, Charles Dela Cruz, Shelli Farhadian, Albert

I. Ko, Saad B. Omer & Akiko Iwasaki. 2020. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* 584. 463–469.

Matsui, Tomoko, Hannes Rakoczy, Yui Mirua and Michael Tomasello. 2009. Understanding of speaker certainty and false-belief reasoning: a comparison of Japanese and German preschoolers. *Developmental Science* 12(4). 602–613.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.). *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. 3148–3163. Reykjavik: European Language Resources Association (ELRA).

Mithun, Marianne. 2025. The Mighty Demonstrative. *Linguistic Typology at the Crossroads* 5-2. 104-122.

Noonan, Michael. 1985. Complementation. In Timothy Shopen (ed.), *Language typology and syntactic description, Vol. II, Complex Constructions*, 42–140. Cambridge: Cambridge University Press.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Accessed 2025.12.28).

Reesink, Ger P. 1993. 'Inner speech' in Papuan languages. *Language and Linguistics in Melanesia* 24. 217–225.

Rumsey, Alan, John Mansfield & Nicholas Evans. 2022. The sound of one quotation mark: Quoted speech in Indigenous Australian narrative. In Alexandra Aikhenvald, Robert Bradshaw, Luca Ciucci & Pema Wangdi (eds.), *Celebrating Indigenous Voices. Legends and Narratives in Languages of the Tropics*, 33–72. Berlin: De Gruyter Mouton.

Saito, Hiroaki. 2021. Grammaticalization as decategorialization. *Journal of Historical Syntax* 5(10). 1–24.

San Roque, Lila, Alan Rumsey, Lauren Gawne, Stef Spronck, Darja Hoenigman, Alice Carroll, Julia Miller & Nicholas Evans. 2012. Getting the story straight: language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation* 6. 134–173.

Sauerland, Uli, Bart Hollebrandse & František Kratochvil. 2020. When hypotaxis looks like parataxis: embedding and complementizer agreement in Teiwa. *Glossa:*

a journal of general linguistics 5(1). 89.

Schnell, Stefan, Geoffrey Haig & Frank Seifart. 2021. The role of language documentation in corpus-based typology. In Geoffrey Haig, Stefan Schnell & Frank Seifart (eds.), *Doing corpus-based typology with spoken language data: State of the art*, 1–28. Honolulu: University of Hawai'i Press.

Schnell, Stefan & Nils Norman Schiborr. 2022. Crosslinguistic Corpus Studies in Linguistic Typology. *Annual Review of Linguistics* 8(1). 171–191.

Spronck, Stef & Daniela Casartelli. 2021. In a manner of speaking: How reported speech may have shaped grammar. *Frontiers in Communication* 6. 624486.

Widmer, Manuel, Sandra Auderset, Johanna Nichols, Paul Widmer & Balthasar Bickel. 2017. NP recursion over time: Evidence from Indo-European. *Language* 93. 799–826.

Corpora and databases

A culturally informed corpus of Dalabon

Ponsonnet, Maïa. 2013. *A culturally informed corpus of Dalabon*. Endangered Language Archive. https://www.elararchive.org/dk0071/ (Accessed 2025.12.28).

Datenbank für Gesprochenes Deutsch

IDS, *Datenbank für Gesprochenes Deutsch (DGD)* [PF_E_00134_SE_01_T_01, FOLK_E_00337_SE_01_T_01, FOLK_E_00337_SE_01_T_02, FOLK_E_00337_SE_01_T_03, FOLK_E_00144_SE_01_T_01, PF_E_00016_SE_01_T_01, ZW_E_00979_SE_01_T_01]. http://dgd.ids-mannheim.de (Accessed 2025.12.28).

SCOPIC Corpus

Barth, Danielle & Nicholas Evans. 2024. *SCOPIC* 1.0 corpus files. SocCog-corp01 at catalog.paradisec.org.au. https://dx.doi.org/10.26278/1YH7-J821.

Sydney Speaks Corpus

Travis, Catherine E., James Grama, Simon Gonzalez, Benjamin Purser and Cale Johnstone. 2023. *Sydney Speaks Corpus*. ARC Centre of Excellence for the Dynamics of Language, Australian National University. https://dx.doi.org/10.25911/m03c-yz22.

The Yurakaré archive

van Gijn, Rik, Vincent Hirtzel, Sonja Gipper & Jeremías Ballivián Torrico. 2011. *The Yurakaré archive*. Online language documentation, DoBeS Archive, MPI Nijmegen. https://hdl.handle.net/1839/00-0000-0000-0016-662E-4.

Contact

nicholas.evans@anu.edu.au