1

Language Documentation & Conservation Special Publication No. 12
Social Cognition Parallax Interview Corpus (SCOPIC)
ed. by Danielle Barth and Nicholas Evans, pp. 1–21
http://nflrc.hawaii.edu/ldc/sp12
http://hdl.handle.net/10125/24742

SCOPIC Design and Overview

Danielle Barth and Nicholas Evans

Australian National University

and the

ARC Centre of Excellence
for the Dynamics of Language

This paper provides an overview of the design and motivation for creating the Social Cognition Parallax Interview Corpus (SCOPIC), an open-ended, accessible corpus that balances the need for language-specific annotation with typologically-calibrated markup. SCOPIC provides richly annotated data, focusing on functional categories relevant to social cognition, the social and psychological facts that place people and others within an interconnected social context and allow people to interact with one another. By 'parallax corpus' we mean 'broadly comparable formulations resulting from a comparable task', to avoid the implications of 'parallel corpus' that there will be exact semantic equivalence across languages.

We describe the data structure of the corpus and the language functions being annotated, and provide an example of a typological analysis using recursive partitioning, a modern statistical technique.

The current paper should be seen as the introductory chapter of an open-ended special issue of LDC whose goal is to make available both the original corpus, the evolving annotated versions, and analyses coming from them, so that any investigator can examine the corpus with their own questions in mind. A range of new papers, linked to the evolving corpus, will be added to this special issue over time.

1 Introduction. The Social Cognition Parallax Interview Corpus (SCOPIC) provides naturalistic but cross-linguistically-matched corpus data with enriched annotations of grammatical categories relevant to social cognition. By 'parallax corpus' we mean '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 problem with that, from a semantic typologist's point of view, is that it can only be achieved by privileging the semantic structure of the source language in the translations, and that it prevents us from studying the fundamental question of how languages—or the formulation practices of speech communities—bias the expression of particular categories in language-specific ways. The English term parallax, popularised in linguistics through Paul Friedrich's influential book *The Language Parallax*, is defined more generally as 'a change in the apparent position of

an object relative to more distant objects, caused by a change in the observer's line of sight toward the object' (http://www.thefreedictionary.com/Parallax). Here it is primarily either the structures of the particular languages represented in the corpus, or the discourse practices of the speech communities represented by them, which cause the change in how phenomena relevant to social cognition get formulated.

Development of the corpus is part of a project[1] using an innovative stimulus-elicitation methodology in a language documentation context that results in exposition, cooperative conversation and narrative data in different task phases (San Roque et al. 2012). This methodology innovates in four ways:

(a) like many elicitation protocols it aims to increase the density of corpus attestation in a particular semantic domain; in this case it targets the broad spectrum of categories needed for social cognition, a field that has not previously been examined in an integrated way,

(b) by allowing speakers to choose their own formulations for the same situations, 'first-text-bias' effects are eliminated, that is there is no bias coming from the original language of elicitation in what categories are expressed,

(c) the card-sort format makes the protocol a narrative problem-solving task, eliciting high levels of speaker involvement, and

(d) the four task-phases (individual picture description, narrative problem-solving, third-person narrative, first-person narrative) induce different formulations and language choices revealing important dynamic variation in coding within each language, in addition to the cross-linguistic variation presented by the whole corpus.

This task has already been used by a number of linguists, resulting in a rich set of data for typological comparison, and the number of languages for which data is recorded is steadily growing.[2] From a subset of 24 of the over thirty languages in which this task was run, an annotated corpus has been created; we are steadily increasing both the number of languages and the depth and quality of annotation. In this corpus, a common set of annotations are used to identify and label instances of language use that relate to social cognition, so as to provide enriched data for typological analysis. Metadata describing the particularities of the annotations for each language provide

---

[1] This work originated as a project funded by the Australian Research Council (Language and Social Cognition: The Design Resources of Grammatical Diversity; DP0878126), when the actual task was developed and many recordings made. Continuation into the present phase of analysis, in particular the employment of Barth as a postdoc and the funding of the ongoing workshops bringing together language-specific investigators for 'annotation jams', has been made possible by an Anneliese-Maier Forschungspreis awarded to Evans by the Alexander von Humboldt Foundation and the German Federal Ministry of Education and Research, plus support from the ARC Research Centre for the Dynamics of Language (CoEDL), funded by the Australian Research Council (CE140100041). We thank the above-named institutions for their generous support of our research. We also thank Susan Ford for her meticulous job in formatting the manuscript, as well as Dylan Evans for producing the graphic in Figure 2.

[2] Languages include Awiakay, Bulgarian, Burmese, Iwaidja, Momu, Raga, Spanish, Ungarinjin, and Yolhmo, in addition to the languages that feature in SCOPIC.

language-specific descriptions of use and scaffold both language-specific and typological analysis.

This article provides an overview of the structure and purpose of SCOPIC, including its component parts and rationale for the particular functional categories that have been analysed in each language. It forms the entry point to an evolving series of articles, some language-specific and some cross-linguistic, which link to SCOPIC and enhance its usefulness, interpretability and interrogatability for all interested users.

1.1 Research Context. Social cognition is what allows individuals to interact with one another (Frith & Frith 2007). Speakers have knowledge of social facts (e.g. kinship, status, ownership) that place themselves and others within an interconnected social context, and of psychological facts about their own feelings, attentions, desires and their estimations of these for others (San Roque et al. 2012). Elements of social cognition are encoded in many parts of a language's expressive resources including morphosyntax, lexis, prosody and gesture, though our primary focus here is on morphosyntax.

Not all corpora will contain linguistic expressions relevant to social cognition to the same degree. For example, a procedural recipe or a narrative by a lone traveller in the landscape is likely to score low on relevant expressions, whereas a soap opera-style corpus will score high. The Family Problems Picture Task was designed to encourage the expression of social-cognition relevant categories in a number of ways. First, it includes depictions of socially-pregnant and emotionally-charged situations. Second, it elicits naturalistic interaction between speakers (and later, an audience) as they solve a narrative problem. Third, it induces different packaging for the same events as between third-person and first-person narratives. And fourth, it gets participants to return several times, in a natural way, to the characterisation of the same events, giving them the opportunity to exhibit alternative ways of depicting the same thing.

The Family Problems Picture Task is organized around 16 cards that participants must describe and organize into a narrative. A selection of the cards, developed by Alice Carroll in consultation with Evans, Alan Rumsey and Darja Hoenigman, is shown in Figure 1.

Participants work through four task phases. Phases 1 (description) and 2 (problem-solving) involve pairs of speakers; phases 3 (third-person narrative) and 4 (first-person narrative) bring in an additional person as audience. Since this person was not there for the first two phases, their arrival is intended to bring out aspects of audience design and re-evaluation of common ground as the first two participants reframe their narrative.

In Phase 1 the initial pair of speakers describe each card, one at a time, at whatever level of detail they feel is warranted. The cards are not presented in an order that has any logic from a narrative point of view and it is not initially obvious that the same cast of characters is involved.

Figure 1. Four Pictures from the Family Problems Picture Task: (a) Homecoming (b) Release from gaol (c) Drunken gossip (d) Imagining return

In the problem-solving Phase 2 the participants have to sort the cards into a coherent narrative order. They are explicitly told that the actual order is up to them, and that there is no 'right' answer. During this phase one typically encounters many directives (put this card here!), questions (do you think this guy is the same as the man in this other card?) and argumentative moves (this must come before this one because…). The cards were designed to make this far from simple, including potentially conflicting clues, so as to engender maximal discussion by the participants. They also must pay close attention to such issues as the emotions expressed by the characters, e.g. the contrast between the joyous reception a man imagines himself receiving when returning home from jail (Figure 1d), and the frosty reception he actually receives (Figure 1a), owing to earlier violence against his wife.

In Phase 3, a third person who was not present for the first two phases is brought in as an audience and the first two participants are asked to narrate the story they have constructed; it is left to them whether to do this in tandem, or to choose one to do it. In Phase 4, we ask for a first-person narrative from the point of view of one of the characters in the story; the speaker is free to choose which character to identify with.

Though there are naturally variations in the coherence and quality of the material, a striking number of recordings exhibit virtuosic and moving renditions of the narrative. In many situations we have recorded, the fact that the task is clearly fictitious, rather than a traditional story, frees speakers up to be creative rather than face social censure for not correctly rendering a traditional narrative.[3]

We note three other differences between the four task phases which are relevant to linguistic choices. Firstly, the first two phases (especially the second) are designed to be interactive between the two participants, so that a range of footing and speech-act issues, such as choice of address terms between participants or formulation of suggestions, arise from the interpersonal dynamics of the task phase. Once we reach phases three and four, and the speaker(s) are primarily addressing an audience, there are often marked changes in formulation accompanying the shift from conversational to narrative style. Secondly, there are marked differences in the degree of self-consciousness between different task phases: typically speakers choose their words fairly carefully in Phase 1, abandon conscious monitoring while engrossed in the problem-solving task of Phase 2, then return to a more consciously eloquent register in Phases 3 and 4. Thirdly, there are epistemic differences that can play an important role in formulation—for example the social relations between two characters (say a woman and an old man) may not be 'known' when they first appear together on a card in Phase 1, but by the time of Phases 3 and 4 an imaginary world has been constructed in which the relationship has become a given (e.g. the old man is the woman's father). These differences can impact on such choices as whether to formulate reference though a life-stage term without relationality (e.g. the old man) or a kin term (e.g. her father).

The task design thus gives rise to a three-dimensional set of possible comparisons in how a particular scene is described:

(a) comparisons which hold speaker and language constant, but vary the task phase, e.g. linking descriptions of the 'homecoming' scene as between Phases 1, 2, 3 and 4

(b) comparisons between how different speakers of the same language describe the same scene

(c) cross-linguistic comparisons of how the same scene is described in different languages comprising performances by different groups of speakers of the same language

A central design principle of SCOPIC is that it allows comparison on any, some, or all of these dimensions for a given scene.

A second key feature of our corpus is that our annotations are organised along functional categories. Each language in the study is annotated for expressions that relate to many functional categories relevant to social cognition. These annotations are organised around the principle of coding a cross-linguistic category, and then the language-specific instance of that category. Within each broad functional category,

[3] This is not to say that no researchers administering the task faced problems. Some researchers had difficulty getting participants to understand the task, or the visual conventions in the pictures such as thought bubbles, or to be motivated to construct a coherent narrative. Roughly 20% of task administrations faced problems of this nature. In our decisions about which languages to include for detailed analysis we have left out those which did not yield rich results.

researchers code a "TAG" and a "TERM" for each instance of the phenomenon. A TAG comes from a closed and cross-linguistically fixed list of category choices and indicates the type of expression being used for the relevant instantiation of a particular functional category. A TERM is the citation form of a language-specific instance of that phenomenon. The same tag, e.g. QPFM (quotation framed by speech element) can be used with many different language-specific terms, such as QPFM_say, QPFM_tell, QPFM_ask, QPFM_talk where these are all instances of quotative predicates that frame some kind of quoted speech, but using different verbs to do so. The same TERM (i.e. linguistic form) may also sometimes appear in different tagged categories, as in: QPFM_talk and QPM_talk where the verb *talk* is a quotative predicate that may in one instance frame an element of quoted speech element (QPFM) but in another instance be used without any specified quoted speech element (coded QPM), as in "they were talking 'let's go!'" (QPFM) v. "they talked together" (QPM). This reflects the widespread many-to-many mapping between function and form in language.

A corpus-based typology project such as SCOPIC allows us to explore the variability within functional domains both inside a language and across languages. With enough data, we are able to investigate variability within the speech of participants as well. With the use of multifactorial statistical techniques, we can ask not only which languages do what to what extent, but what internal and external predictors have an effect on all or some languages and why this may be the case.

1.2 Languages. SCOPIC currently has media and data from 24 languages; it is expected that there will be additions from other languages over the course of the project, and additional data added from each of the languages currently listed. Table 1 shows the languages that are part of the project and the researchers who have collected the media and annotated the data. As is clear from the table below, the languages in the project come from all over the world, with representation from every continent and from many different language families, though with a slight bias towards Oceania given our geographical base. Languages were selected for the project based on the basis of having grammatical characteristics of known importance for the grammar of social cognition, supplemented by two creoles (North Australian Kriol and Bislama from Vanuatu) chosen for their potential relevance to illustrating processes of rapid grammaticalisation replicating categories in one or more substrate/adstrate languages in the sample. However, all languages, and all speakers, construct social relationships and negotiate the relevant information by some linguistic means, so the functional categories that were examined in the project are in principle relevant to any language.

Table 1. Languages in SCOPIC

Language	Family	Country	Macroregion*	Researcher(s)
Amharic	Semitic	Ethiopia	Africa	Hirut Woldemariam and Mengistu Amberber
Auslan	Signed Language	Australia	Sign Language	Gabrielle Hodge
Avatime	Niger-Congo	Ghana	Africa	Saskia van Putten
Balinese	Austronesian	Indonesia	Island Southeast Asia	Wayan Arka
Bislama	Creole	Vanuatu	Creole	Stefan Schnell

Language *continued*	**Family**	**Country**	**Macroregion***	**Researcher(s)**
Dalabon	Gunwinyguan, Australian	Australia	Australia	Nicholas Evans
Duna	Trans New Guinea	Papua New Guinea	Pacific/Melanesia	Lila San Roque
English	Indo-European	Australia	Europe	Barbara Kelly, Danielle Barth and TBD
German	Indo-European	Germany	Europe	Andrea Schalley
Hoocąk	Siouan	United States	North America	Iren Hartmann
Idi	Pahoturi River	Papua New Guinea	Pacific/Melanesia	Volker Gast
Japanese	Altaic	Japan	Asia	Eri Kashima, Heiko Narrog and Nicholas Evans
Khalkha Mongolian	Mongolic	Mongolia	Asia	Elena Skribnik and Dolgor Guntsetseg
Kogi	Chibchan	Colombia	South America	Henrik Bergqvist
Komnzo	Yam	Papua New Guinea	Pacific/Melanesia	Christian Döhler
Kriol	Creole	Australia	Creole	Gregory Dickson
Ku Waru	Trans-New Guinea	Papua New Guinea	Pacific/Melanesia	Alan Rumsey
Matukar Panau	Austronesian	Papua New Guinea	Pacific/Melanesia	Danielle Barth
Murrinhpatha	Southern Daly, Australian	Australia	Australia	John Mansfield
Sanzhi	N.E. Caucasian	Russia	Caucasus	Diana Forker
Sherpa	Tibeto-Burman	Nepal	Asia	Barb Kelly
Tok Pisin	Creole	Papua New Guinea	Creole	Danielle Barth and TBD
Vera'a	Austronesian	Vanuatu	Pacific/Melanesia	Stefan Schnell
Yurakaré	Isolate	Bolivia	South America	Sonja Gipper

*With homeland macro-region used for English, and putting signed languages in their own category; Auslan derives historically from British Sign Language.

1.3 Categories relevant to Social Cognition – Functional Categories/ Functional Equivalents. Social cognition is the sum of those psychological processes that allow us to interact and live with each other. This in turn is the key to our being human, and to the very possibility of human society and culture.

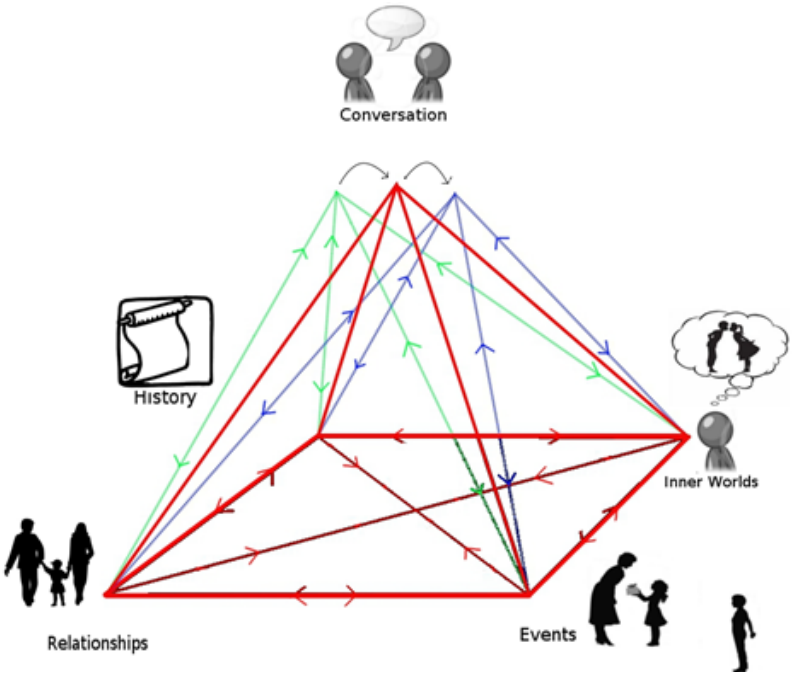

Figure 2. Elements in an overall model of social cognition in language. The advancing peaks of the pyramid symbolise the potential updating of all elements which accompanies each conversational move

Social cognition is woven from many threads. We need to represent what others are wanting, thinking and feeling ('inner worlds' in Figure 2). We need to represent the complex web of relationships that links the actors and objects in our social universe, and model the social mores and expectations that govern and partially predict how they will behave ('relationships'). We need to work out what each action means for all involved, to apportion credit and blame, benefit and obligation, and keep track of whether actors have achieved their goals or are thought to have acted competently ('events'). We keep tallies of how this all changes through time ('history'). Most centrally ('conversation'), we constantly update, through conversation, our own and each other's mental models of our social universe, as well as our, and their, attention, feelings, knowledge and beliefs. And, with almost every move, we bring about changes in what our companions do and think through what we say and how we say it (shown by the advancing peaks of the pyramid).

These elements are interlinked in numerous ways, symbolised by the bidirectional linking lines. To give a single example, the choice between an intimate and a respectful pronoun (*tu* vs *vous*, *du* vs *Sie* etc.) is instantiated in the face-to-face environment of the *conversation* but depends at the very least on *relationships* and may interact with each of the other elements, e.g. history (if our choice of *du* depends on a prior 'breakthrough' in intimacy), inner worlds (e.g. if I assess you as the sort of person who thinks I call you *tu* because I think you will work out that I am doing so, despite your stranger status, because it will be evident to you that we share an egalitarian and informal political ideology), and events (e.g. where the depicted event—say a favour sought of the addressee—is more likely to lead to politeness upgrades).

The encoding of all this in speech is achieved through subtle and complex linguistic choices linked together in conversational turns whose timing is measured in microseconds. Despite the speed and complexity of the task, the social stakes can be extraordinarily high: in extreme cases a single false move can lead to ridicule, disgrace, divorce, or worse.

Each of the major architectural elements above can further be broken down into numerous parts. Conversation includes, inter alia, the alternation of speaker and addressee roles, the deictic field, the negotiation of common ground, and the conveying of appropriate social footing through appropriate address terms. Comparable expansions could be made for each other element. Investigating each of these elements is a substantial project, and we do not tackle them all at this stage. For the purposes of this phase of engaging with the corpus, we focus on:

(1) Reference to Humans (rubric: 'relationships')

(2) Expressions of Reported Speech and Thought (rubric: 'inner worlds')

(3) Depictions of Social Ramifications (rubric: 'events')

(4) Evidentiality, Stance, Perception and Evaluation (as an example of interaction between, at least, 'conversation', 'events', and 'inner worlds'.

Over the course of the project, additional functional categories will be added, but for the time being we concentrate our discussion on 1–4 above. Each of these functional categories have annotations in one (or more) separate tiers (cf. §2. Corpus Structure –Theoretical Organization).

1.3.1 Reference to Humans. The Family Problems Picture Task Cards depicts different configurations of characters engaged in different social situations. Participants can describe the characters in many ways: generically (*a man*), as a kinship term (*the father*), a kinship term in relation to someone else (*his father, his friend*), descriptively *(the mad one, the tall one*), with a pronoun (*he*), with a pointing gesture, or in many other ways. As this list indicates, some methods (e.g. kin terms) are relational, others are not (e.g. the tall one). Speakers have the choice, in formulating each reference to persons (Enfield & Stivers 2007), in how each referent is characterised, and this choice gives them the chance to highlight, ignore, seek information about, or redefine social relationships.

We are interested here in the differences in how these decisions in formulation are made—whether across individuals, task phases, or languages. Do you use a kinship term in a family scene but a generic word in a scene depicting violence? Do you use a possessed kinship term when you are telling a narrative in first person more often than when you are telling a narrative in third person, or negotiating the order of the cards to make up a narrative? How consistent are the favoured strategies across different speakers of a language, and across task phases? Are kinship terms obligatorily possessed in some languages and do these languages then use kinship terms proportionally more or less than other languages? These are some of the questions we can answer by investigating and annotating human referents in the corpus.

1.3.2 Expressions of Reported Speech and Thought. People engaged in the Family Problems Picture Task describe the thoughts, motivations, speech and emotions of both the characters and themselves (at a meta-discourse level). We are interested in how participants talk about talking, thinking and emotions. We are particularly in-

terested in how participants report on the mental states and discourse of themselves and others. Participants may enact the speech or thought of someone directly, or they may frame a reported speech or thought with a verb of speech, emotion, cognition, or perception or something else. Through annotating these kinds of reports of discourse and mental states and how they are framed, we can answer questions about the variety of expressions in each language and what kinds of strategies predominate.

1.3.3 Event Depictions: Social Ramifications. The Family Problems Picture Task cards depict characters engaged in different social situations. Participants can express how these events express people and how. Take the famous event in Australian history when Prime Minister Gough Whitlam poured sand into the hand of Gurindji activist Vincent Lingiari (Figure 3). This could be described simply as a physical event: one man pours sand into the hand of another. But its social ramifications were huge: the historical entitlement of Aboriginal people to their land was being recognised through this act, with repercussions that are still being played out today. Less visible from the photo is another aspect of the event, namely the intention of the participants: Whitlam knew that Lingiari was blind, and by pouring sand chose a means of communication that was tactile rather than visual—in other words, his motives drew on knowledge of what his interlocutor could readily attend to.[4] In many events, then, on top of a 'physical component' (e.g. movement of objects through space, spatial layouts) there is overlaid a social component with all sorts of consequences for such matters as future obligation, changes to who knows what, or the reconfiguring of social relationships. And obviously social roles—such as Prime Minister or indigenous activist/tribal leader—are also deeply relevant.

In more general terms, event depictions might express who benefits (benefactive constructions), who suffers (malefactive constructions), whether actions were jointly undertaken (comitative, reciprocal and assistive constructions), intentions of action (volitional, apprehensive, intentionality constructions), whether the action was known to others, and if those actions led or did not lead to the result that the agent had planned (frustrative constructions). Languages also differ to what extent speakers must clarify whether or not they have the epistemic authority to estimate a person's inner-world or state, at what kind of private predicates speakers can attribute to someone. These kinds of expressions, among others, are annotated in the corpus to investigate the range of devices languages have to express social ramifications of actions, and how often these expressions are used.

[4] As witness Kev Carmody said: "You'll notice Gough pulls the old man's hand up because he can't see too well. The old man … he doesn't smile when Gough lifts his hand up. As soon as that sand hits his hand, you can see he grinned. 'Got'em!'" (Interview on ABC George Negus Tonight 'The Gurindji Strike' 5/7/2004). We are grateful to Felicity Meakins for drawing this aspect of the event to our attention.

Figure 3. Prime Minister Gough Whitlam pours soil into the hands of traditional land owner Vincent Lingiari, Kalkarindji (Wave Hill), Northern Territory, Aug 16, 1975
Printed 1999 Art Gallery of New South Wales Hallmark Cards Australian Photography Collection Fund 1991 Photo: AGNSW © Mervyn Bishop (Australia, b.1945) / Department of the Prime Minister and Cabinet.

1.3.4 Evidentiality, Stance, Perception and Evaluation. The Family Problems Picture Task cards deliberately include ambiguous situations and situations in which the depicted characters have incorrect presuppositions about events. We are interested in how participants report on characters', and their own, confidence with regard to what is happening. Which languages have grammaticalised strategies for encoding uncertainty and which do not? Does having relevant grammaticalised categories in particular languages impact on the frequency of occurrence of such expressions of evaluation due to codified categories in those languages? How do task participants express the knowledge, and knowledge source, imputable to them and to their characters?

2 Corpus Structure – Theoretical Organization. Table 2 summarises the corpus components and a further discussion of those components follows in the sections below.

Table 2. Primary and supplementary files in the corpus

a) Per Language	**Format(s)**	**File Name**
File Metadata	CSV	SocCog-ISO00-sessions.csv
Speaker Metadata	CSV	SocCog-ISO00-speakers.csv
Annotation Metadata (1 for each variable)	CSV	SocCog-ISO00-tag_term_ variable_key.csv
b) Per Session		
Video recording(s) (1-4 parts)	MPEG, MOV, MP4	SocCog-ISO00-Title.extension
Audio recording(s) (1-4 parts)	WAV	SocCog-ISO00-Title.wav
Transcription, Translation and Annotation (1-4 parts)	EAF	SocCog-ISO00-Title.eaf
c) For Entire Corpus		
Closed Vocabulary Descriptions (1 for each variable)	DOC	SocCog-SCOPIC00-Title.doc
Datasets (1 for each variable, 1 with all variables)	CSV	SocCog-SCOPIC10-Title.csv
Card Descriptions	CSV	SocCog-SCOPIC20-Title.csv
R scripts (corpus compiling, data-set production, frequency counts, graphing, example analyses)	TXT	SocCog-SCOPIC30-Title.txt

2.1 Corpus Components. A corpus is simply a body of text that is computer readable and structured in some way. The SCOPIC corpus includes accompanying media files, both audio and audio-visual. A corpus must also have metadata, so that the information in it is interpretable. At the core of SCOPIC is a series of ELAN (.EAF) files, a kind of xml structured text. These files are produced through the linguistic annotation software ELAN (EUDICO Linguistic Annotator) which is a multimedia annotator from the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands (cf. Sloetjes & Wittenburg 2008). ELAN is freely downloadable at http://tla.mpi.nl/tools/tla-tools/elan/. The ELAN files provide transcriptions, translations and annotation of data in media files and are linked to those associated media. ELAN has become a common way to organize data in cross-linguistic corpus studies (cf. Schiborr 2016). The ELAN files have the same referring collection ID (i.e. MJK01, DAL02, AVN03, etc.) as the media files that they are associated with. ELAN files can have multiple tiers for annotation, which can be assigned different "Type Names" and can also be linked hierarchically. The tiers, types, and their organization are described in §2.2 ELAN Tiers.

SCOPIC contains several kinds of metadata and other resources, to make it easily useable by researchers involved in the project and other interested parties. We have metadata at the level of each language: information about the files in the corpus, information about the speakers in the corpus (to the extent permitted by those speakers)

and information about the kinds of annotations for each variable. We also provide an account of the annotation decisions we have made, which detail the categories and aid reproducibility.SCOPIC also contains derived datasets relating to the whole corpus and to each variable of interest. These provide a shortcut for those interested in further study of the corpus data in the form of an exported and cleaned up set of the data; they will be of use to corpus linguists unfamiliar with ELAN and its exporting options. Finally, we provide some sample *R* scripts for data wrangling and doing basic analyses, in the interest of accountability, reproducibility and as an example of some of the uses of SCOPIC.

SCOPIC currently exists in an offline version, but will be available online by the end of 2017. The first online, open-access version of the corpus will contain only the annotated files that researchers, and the communities they work with, feel ready to release. As richer annotations are developed, new releases of the corpus will be available, along with the archived versions. Currently, many media files associated with SCOPIC are available on PARADISEC, a digital archive of materials primarily from endangered cultures at (http://catalog.paradisec.org.au/collections/SocCog). However, visitors to the collection should be aware that this is a collection of audio and audio-visual data of people engaging in the SCOPIC task, it is not yet the SCOPIC corpus. SCOPIC, as a corpus, will be updated annually and will be available at LAC, Language Archive Cologne (https://lac.uni-koeln.de/en/) alongside associated media files.

2.2 ELAN Tiers.

Each ELAN file has several tiers required for the project, as described below. However, researchers in the project are allowed to have variable implementation of these tiers, as best fits their data. At the point where we began to coordinate annotations, many researchers in our project had already transcribed and translated their data, and several had already done interlinearisation. Rather than requiring that researchers either eliminate useful work by deleting tiers, or overburdening them by requiring that everyone uses exactly the same format, we chose to add several systematised tiers to the structure a researcher already had in place. The result of this is that there are particularities for each language/researcher, and all researchers' data fits within their own research programs. More generally, through this policy we hope to steer between the extremes of overstandardisation (insensitive to language- and investigator-specific needs) and overparticularisation (making comparison impossible). Through the systematised social cognition tiers, which each have their own "Type Name", we can export the data relevant to our project. We encourage corpus users interested in particular languages to engage with the ELAN files directly to see the annotations in context.

2.2.1 Transcription.

All ELAN files have a transcription of the data. This primary expression of the data is done by researchers as fits their own research program. Generally segmentation is at the utterance level. Transcription is generally done in the practical orthography of the language or a transliteration of that orthography. In some cases (e.g. Japanese, Amharic) there are separate tiers for the language's standard writing system, and for a romanised transliteration. Each speaker in the file has their own tier in the ELAN file. The type of this tier is: `ts`.

2.2.2 Translation into Research Language.

Translation is generally done into English, but for some languages translation is done into another more widely known

local language (Tok Pisin) or appropriate research language (Spanish, Russian). For each translation tier, there is a dependent translation tier. The type of this tier is: `tl`. There may be multiple dependent translation tiers for each transcription tier (i.e., both Spanish and English).

2.2.3 Multiple Tiers for each Variable of Interest. The main part of our project is the annotation of functional categories relevant to social cognition. For each variable of interest there is an annotation tier with a special type relevant to that category (Human Referent, Reported Speech, PrivPred, Benefactive,[5] and Evaluation) and a notes tier with the type Variable Notes. The category annotation tiers are parent tiers (not dependent on another tier) and the notes tiers are dependent on the variable with which they are associated. Some annotators have chosen to have one variable tier per file, and others have chosen to have a variable tier for each of the participants represented in the file (usually 2 or 3). For annotators who have chosen the latter, there is an indication of which tier belongs to which participant. For those who have chosen the former, the parallel time-stamps with the reference tier (see §2.2.6. Speaker-based Utterance Reference below) provides this information.

2.2.4 Card Tier. The Family Problems Picture Task is organized around 16 cards that participants must describe and organize into a narrative (see San Roque et al. 2012 for fuller description). There is a tier with the type FPPT_CARDS for each file. This tier uses a closed vocabulary of card numbers and standardised labels (i.e. 1_Homecoming or 2_Receiving clothes) to indicate which card participants are discussing or referring to at each moment of discourse. This may be indicated by several (often 16) long annotations, or by many short annotations with time stamps that match each moment of transcribed discourse. Having this tier allows us to investigate whether there are certain cards (or kinds of cards) that generate particular kinds of discourse cross-linguistically, and also to compare different descriptions of the same card by the same speaker(s) at different task phases.

2.2.5 Additional Structural Information (Optional/Intermittent). For some languages in the corpus, the data are fully interlinearised, with parsing and glossing according to the Leipzig Glossing Rules (Comrie et al. 2008). For most languages, however, parsing and glossing is only intermittent and confined to the cases where it is necessary. This additional structural information may be used when an annotation of a variable of interest needs more clarification, such as a verb with a special affix or a multi-part construction used to implement a functional category. This information is across 2 tiers (one for parsing, one for glossing), where the first parsing tier is dependent on the relevant transcription tier and has the type `morph`. The second glossing tier is dependent on the parsing tier and has the type `gloss`.

2.2.6 Speaker-based Utterance Reference. For each transcribed utterance, there is a unique reference number that indicates the language, file, speaker and order of their utterance. There is a tier for each speaker. The type of these tiers is: `ref`. They are parent (non-dependent tiers).

[5] PrivPred stands for Private Predicates. Both Private Predicates and Benefactives are types of Social Ramifications of Depicted Events.

2.2.7 Other Tiers. Particular annotators may have other tiers in their files that are relevant to their individual research programs. These can provide more information if examined within context. For cross-linguistic analyses, we recommend exporting information from ELAN based on types, so that only the information of interest is compiled.

2.3 Metadata. There are several kinds of metadata for the project, produced for each language, each functional category of interest and for the corpus as a whole. The metadata is generally in a structured but simple text format (.csv) so that it can easily be merged into the corpus using scripts in R or other data wrangling strategies like VLOOKUP in Microsoft Excel. Scripts for accomplishing this are also provided (cf. §2.5. Scripts for Application of the Corpus).

2.3.1 Metadata for Functional Categories

2.3.1.1 For the project. For each functional category, we have a closed set of annotation choices developed by the project researchers. In general these choices arose in a bottom-up way from discussions resulting from hands-on workshops over several days. These involved researchers for all the languages listed in Table 1, split across two meetings per year, one in Europe and one in Australia, with researchers divided between these according to the practicalities of travel.

These sets of annotation choices are meant to provide a balance between language-specific features arising from the project language sample, and general cross-linguistic comparability. Documentation is provided which outlines the different coding choices and how to determine whether a particular language usage falls into that category. In the parlance of the project, these annotations are called TAGs, which combine with TERMs. Language-specific implementations of these annotation choices were done by language experts, noting a TAG (general category) and then a TERM, which is a language-specific instance of the TAG. Citation forms of words were used as TERMs unless there was a specific reason not to. So, for example, for a particular verb of speech, an annotator might use the TAG EPF (for 'cognitive predicate that frames speech') and the TERM *think* as the citation form for instances of *think, thinks* or *thought*, since the tense and person inflection of the predicate is not necessarily relevant to speakers' construction of mental states or reported discourse. However, with the TAG PKN (possessed kinship term), an annotator would indicate the possessor for *my aunt* and *her aunt* respectively in the TERMS as *aunt.1s* and *aunt.3s* because the person of the possessor matters for the construction of relationships.

Because there is a lot of instructional information, examples and description, this metadata is provided in a .doc format.

2.3.1.2 For each language. For each language in the project, metadata is provided for each unique instance of a TAG_TERM combination. An annotator provides a .csv document with each TAG_TERM in a row in the file (for each functional category), an English translation of that TAG_TERM and relevant information about it. The list of unique TAG_TERMs is automatically generated from the ELAN files using an Export As.List of Words function in the program.

2.3.2 Metadata for Sessions. Each run of the Family Problems Picture Task is called a session. Most sessions have 3-4 parts (Card description, Negotiation of cards into a

narrative, Telling the narrative in third and first person) but some have less and some have more, depending on the particular language and the participants from that task. A metadata sheet is provided by each annotator in .csv format with information about each session including who the participants were, which parts of the task were done, where it took place and so on.

2.3.3 Metadata for Participants. Each annotator has provided a .csv file with information about the participants who did the Family Problems Picture Task which includes their names (unless they preferred not to provide this information), approximate ages, gender, language background and so on.

2.4 Derived Datasets. In the process of analysing patterns in our corpus, we have created several datasets: one for the entire corpus, and one for each separate variable of interest/functional category. For the sake of replicability, transparency and ease of use by researchers outside the project, these datasets are available to outside users of the corpus. As the project evolves, new data will be coded and new datasets need to be generated to reflect the current state of research. Because of this, datasets are versioned. We recommend that researchers use the most current version of the datasets.

2.5 Scripts for Application of the Corpus. In the process of preparing data for the analysis of typological patterns in the corpus, we have developed several computer scripts in the R programming language (R Core Team, 2015). The format of the scripts is plain text (.txt files). These scripts are for data wrangling (organising the data into an analysable format), for exploration, and for generating descriptive statistics. We also have several scripts for inferential statistical analyses. A selection of these scripts are provided with the corpus. As the project develops, new scripts will be added and new versions of the included scripts will be updated. Updates are versioned. We recommend that researchers use the most current versions of the scripts.

3 Corpus Analyses. SCOPIC is made up of rich, naturalistic, interactive data and there are many potential veins of research that can come of out analysing it. Because of the added value that comes from integrating analyses with the corpus, we plan for these articles to appear in the dynamic special volume of LDC of which this article is the first element, added to the volume as they become available, and linked to the corpus which forms the spine of the research.

Below are some analyses planned by the research group that has produced SCOPIC. Example sentences will ideally link back to the corpus, providing further information in context. A link will point to an ELAN file linked to its media file, both housed in PARADISEC, a digital archive of materials primarily from endangered cultures. Time stamps will be used to point to a specific part of the file. Being able to see and hear examples in context is particularly important for investigations of social cognition where gesture, enactment and non-verbal inter-speaker social cues are used to help express language relevant to social cognition.

3.1 Social Cognition Sketches from Particular Languages. One type of analysis associated with the SCOPIC project will be language sketches for many of the languages represented in the corpus. These will describe how social cognition works in the grammar, giving an overview of a wide range of relevant grammatical phenomena, and will be written by a language expert/experts. These language sketches will

combine the researchers' general data (i.e. from all known work on the language) with observations from SCOPIC data, and will discuss features and topics most relevant for that particular language in relation to social cognition. These articles will come out on a rolling basis, but we expect to have 3-4 finished, edited and ready to publish by the end of 2017. Some example articles will be Social Cognition in Dalabon grammar (Evans), Social Cognition in Duna grammar (San Roque), Social Cognition in Kogi grammar (Berqqvist), etc.

3.2 Category/Variable Analyses for Particular Languages and Language groups. A second strain of analyses coming out of the project focus on particular topics in particular languages, also written by experts on the languages. Not every topic will be relevant for each language in the project, so we expect 1–3 topic articles per language and perhaps some articles on a particular topic for small clusters of languages. Because the analyses will be coming out of the Family Problems Picture Task data, there will be links to particular files and linked references to particular utterances. Links will be updated as analyses and annotation files are revised. These articles will come in on a rolling basis, but we expect to have 2–3 ready to publish by the end of 2017 and another 3–4 coming in through 2018. Some example articles would be Social Ramifications of Depicted Events in Matukar Panau (Barth), Reported Speech and Thought in Avatime (van Putten), Evidentiality in Yurakaré (Gipper), Possessed Kin Terms in Australian Languages (Evans, Mansfield and Dickson), Collective Group Terms in languages of Papua New Guinea (San Roque, Gast, Döhler and Barth), etc.

A third strain of research will be multi-authored, large cross-linguistic typological studies using the data from the Family Problems Picture Task focusing on a particular grammatical feature. We expect 4–5 articles on the main features that have been annotated for in the corpus, such as Reported Speech and Thought, Expressions of Human Referents, Stance and Evaluation and Social Ramifications of Depicted Events. These articles will also have examples which link to the primary media and annotation data through PARADISEC.

4 Suggestions for Corpus Use. We suggest further typological investigations of the languages which have already been annotated, using the available datasets. We also suggest researchers interested in particular languages use the annotations in SCOPIC as a model for coding their own data to determine where their language fits into a larger typological pattern.

What follows is an example of an initial typological sketch for a *Reference to Humans* variable. Even with small amounts of data, we can begin to see patterns of significant difference between languages in their strategies governing the kinds of predicates used for discussing speech, thought and emotions. The patterns are most easily seen by presenting descriptive statistics as bar charts and inferential statistics as classification trees (bar charts give a quick feel for the data, classification trees provide significance testing). Binary classification tree analysis divides the data into two sections based on which datapoints are most different from one another based on the given (levels of) independent variables. Then under each branch of the tree, the data will again be divided into two sections based on which datapoints under the branch are most different from one another based on the remaining given (levels of) independent variables. The analysis can further divide data using the remaining levels of an independent variable that has already resulted in a split. This process of partitioning continues recursively until each branch is relatively homogenous. The data under the branch is considered

homogenous enough when there are no significant differences (using a *p* value) between the datapoints, given the independent variables. *p* values are provided for each branch (cf. Hothorn, Hornik & Zeileis 2006).

Each of the analyses discussed here have data from the following languages (shown with ISO code), coded by the researchers listed:

AUS	Auslan	Gabrielle Hodge
AVN	Avatime	Saskia Van Putten
DAL	Dalabon	Nicholas Evans
DAR	Sanzhi	Diana Forker
DEU	German	Andrea Schalley
DUC	Duna	Lila San Roque
JPN	Japanese	Eri Kashima
KHK	Khalkha Mongolian	Dolgor Guntsetseg and Elena Skribnik
MJK	Matukar Panau	Danielle Barth
MWF	Murrinhpatha	John Mansfield
ROP	Kriol	Gregory Dickson
YUZ	Yurakaré	Sonja Gipper

The set of annotations for the reference to humans category contain twenty possible choices for labelling each reference to a human by a participant. For particular research questions, it makes sense to merge those twenty different categories into larger categories, highlighting distinctions of interest in a given analysis. In this case we want to examine further the construction of relationships and the inclusion or exclusion of people in relationships; therefore we examine if different languages make distinctions in their frequency of use of possessed kinship terms (*my mother*), non-possessed kinship terms (*a mother*), and non-kinship terms (*a woman, a nurse,* etc.).

The bar chart in Figure 4 shows the proportion of possessed kinship terms, non-possessed kinship terms and other terms for each language in the corpus. The bar chart shows that Dalabon, Khalkha Mongolian and Matukar Panau all have a relatively high proportion of kin terms, with Dalabon using kin terms for over half of the human referents. Auslan, Avatime, Sanzhi, German and Duna have very few kinship referents in the currently coded data, and the kinship terms they do have tend not to be possessed. Dalabon, Khalkha Mongolian, Matukar Panau and Yurakaré, languages with a high proportion of kinship terms express all or nearly all of those kinship terms as possessed. For Khalkha Mongolian, it may be a grammatical necessity to possess kinship terms. Japanese and Murrinhpatha, on the other hand have many uses of kinship terms, but most of these are unpossessed terms. Kriol has a relatively even split of possessed kinship terms, unpossessed kinship terms and non-kinship terms in the sample. These patterns show that there is some grouping in the kinds of use, but that it is not determined by region.

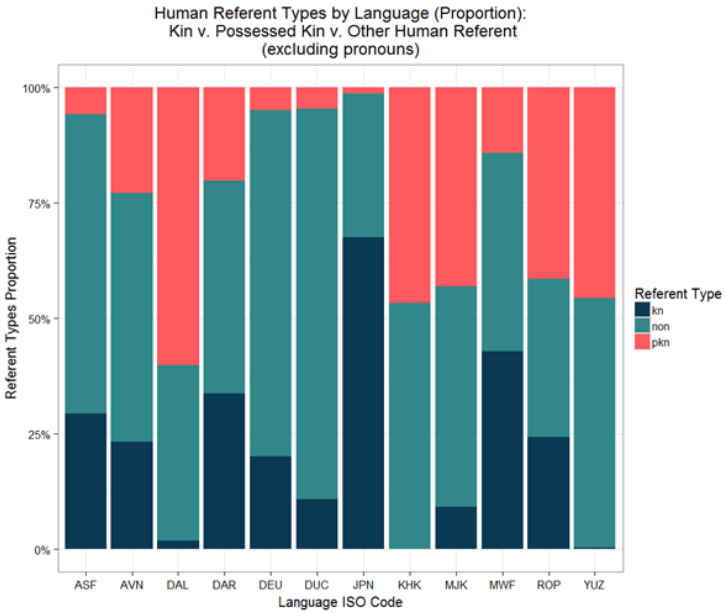

Figure 4. Human Referents Possessed Kinship, Non Possessed Kinship and Other

The classification tree in Figure 5 has language as the sole independent variable. We are interested in where the divisions in language types lie. The figure shows that first there is a significant difference between the languages that have a high proportion of unpossessed kinship terms (left side under node 1) and those that do not. The next split on the left side of the tree is between the languages that have higher proportions of kinship terms than non-kinship terms (under node 2), which are Murrinhpatha and Japanese. These two languages have a further split under node 3 because Japanese has significantly more kinship terms than non-kinship terms. Kriol patterns differently from Auslan, Avatime, Sanzhi and German (under node 6) as it has a fairly even distribution between term types and the other languages have more generic terms. Under node 8, Auslan and German pattern together as having a high proportion of generic terms and few possessed kinship terms, and Avatime and Sanzhi pattern together as having more generic terms, but comparable proportions of possessed and non-possessed kinship terms. Under the left side of node 1, we see the languages that have few non-possessed kinship terms. Duna is the most different in this group, having very few kinship terms overall and primarily using non-kinship terms (under node 11). Under node 13, Matukar Panau patterns against the Dalabon, Khalkha Mongolian and Yurakaré in that it has a moderate amount of kinship terms and most of the kinship terms used by speakers are possessed ones. The split under node 15, where languages have almost necessarily possessed kinship terms, shows that Dalabon has more kinship terms than other kinds of terms. Khalkha Mongolian and Yurakaré have use more non-kinship terms, but where they do use kinship terms, they are highly likely to be possessed.

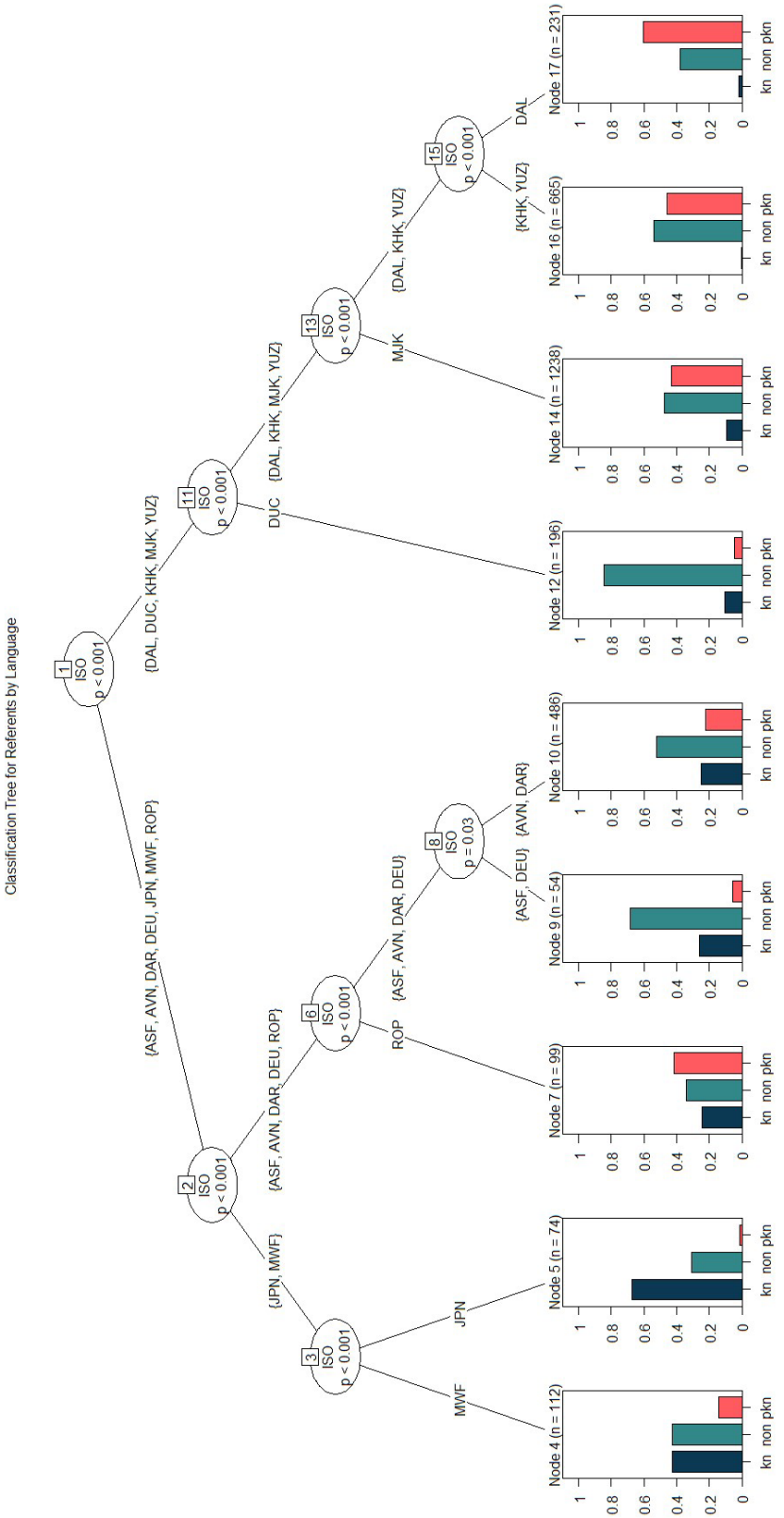

FIGURE 5. Classification Tree for Human Referents Possessed Kinship, Non Possessed Kinship, and Other

Through this data, we see that some languages have a much stronger tendency to use kinship terms for human referents (Dalabon, Japanese, Kriol) while others seem to avoid them (Duna, German, Auslan). Some languages use kinship terms to establish relationships between the characters in these stories, but do not mark them as possessed (Japanese, Murrinhpatha). Other languages have a very strong propensity (and for some languages, likely a grammatical necessity) to possess kinship terms when they are used (Dalabon, Khalkha Mongolian, Matukar Panau, Yurakaré). The kinship relationship must always be connected to someone, further strengthening the ties between the referents. Matukar Panau has a kinship system in transition, where previously all referential kinship terms were possessed, but due to changes in the language this is no longer a grammatical requirement, although it seemingly stays a strong tendency. This time of transition and its grammatical consequence is realized in its patterning above.

Although these data come from only a small portion of our corpus, it is clear that techniques developed for larger datasets from majority languages like English can be applied to small datasets and to non-majority languages. We also see that having a rich set of annotations allows a researcher to frame a question and then divide up the data as they see fit to answer their particular question. With corpus typology data, we can ask not only does a language have X, but how often do speakers of a language use X vs. Y vs. Z when all of these strategies are available to them? Finally, we can distinguish what seems like common or uncommon patterns of distribution and examine whether that pattern exists in other genres of texts in the same languages, and whether the same patterns of distribution exist in other languages.

References

Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Enfield, Nick J. & Tanya Stivers (eds.). 2007. *Person reference in interaction: Linguistic, cultural and social perspectives*. Cambridge: Cambridge University Press.

Friedrich, Paul. 1986. *The language parallax: Linguistic relativism and poetic indeterminacy.* Austin: University of Texas Press.

Frith, Chris D. & Uta Frith. 2007. Social cognition in humans. *Current Biology* 17. R724–R732.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.

R Core Team. 2015. R: A language and environment for statistical computing. https://www.R-project.org/

San Roque, Lila, Lauren Gawne, Darja Hoenigman, Julia C. Miller, Alan Rumsey, Stef Spronck, Alice Carroll & Nicholas Evans. 2012. Getting the story straight: Language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation* 6. 135–174.

Schiborr, Nils N. 2016. Multi-CAST corpus overview and description. In Geoffrey Haig & Stefan Schnell (eds.), *Multi-CAST (Multilingual Corpus of Annotated Spoken Texts.* https://www.uni-bamberg.de/fileadmin/aspra/Multi-CAST_corpus-overview.pdf (Accessed 03-06-2016.)

Sloetjes, Han & Peter Wittenburg. 2008. *Annotation by category – ELAN and ISO DCR. 6th International Conference on Language Resources and Evaluation (LREC 2008)*.