Annotations using GRAID

(Grammatical Relations and Animacy in Discourse)

Manual Version 7.0

Geoffrey Haig
Bamberg University
geoffrey.haig@uni-bamberg.de

Stefan Schnell
La Trobe University
s.schnell@latrobe.edu.au

October 2014

Contents

1 Introduction

The acronym 'GRAID' stands for 'Grammatical Relations and Animacy in Discourse'. GRAID is a system of symbols and conventions for glossing the grammatical relations and overt forms (noun phrases, pronouns etc.) of major clause constituents in texts. The purpose of GRAID annotations is to facilitate cross-corpus research in language typology. This is an updated version of the GRAID manual 6.0 from October 2011[1].

The system was developed on the basis of transcribed recordings from typologically diverse languages, using data that had been collected and archived in language documentation projects (cf. Haig et al. 2011a). The current standard version of the system uses approx. 30 symbols (cf. Section 5 for an overview) and simple conventions for combining them. GRAID is quite flexible and allows different levels of detail for glossing different items. Thus annotators are in a position to create their own solutions to language-specific problems of glossing. Furthermore, provision is made to allow items to remain unclassified. Although we do not claim that the system of categories implemented in GRAID is necessarily valid for all languages, we believe that the vast majority are amenable to analysis in these terms. Ultimately this is an empirical question, which can only be resolved through experience.

Since releasing the 6.0 version of the manual, we have annotated spoken language texts in a number of additional languages, including North Eastern Neo-Aramaic (Semitic), Northern and Central Kurdish (West Iranian), Teop (Oceanic), Cypriot Greek (Hellenic) and English. The experience gained through working on these languages, and the ensuing discussion with colleagues, have prompted a revision of the annotation practice in a number of respects, which are reflected in this updated version of the manual. Most of the changes outlined here are thus empirically motivated, and represent our solutions to challenges that have arisen over the last two years. We have nevertheless retained much of the basic background information from the preceding version of the manual, but have modified certain aspects considerably (cf. Section 1.4 below for a summary of changes).

1.1 Research context and motivations for GRAID annotations

GRAID annotations are intended to serve as a basis for quantitative typological investigations of natural discourse, of the type pioneered in the work of John DuBois (1987, Du Bois et al. 2003), Balthasar Bickel (2003, Stoll and Bickel 2009), and Noonan (2003), among others. In addition to the syntactic function and morphological form, GRAID annotations also register animacy features of referential expressions. Hence, GRAID-annotated text corpora facilitate additional research questions in the area of animacy and referential hierarchies in discourse (cf. Haig and Schnell 2009 and Haig et al. 2011a for an overview of research topics amenable to GRAID annotations).

[1] We are indebted to Dagmar Jung, Ulrike Mosel, Meytal Sandler, Hanna Thiele and Claudia Wegener for their constructive criticism and additional data, which contributed to this version in more ways than we can do justice here. We would also like to thank Nils Schiborr for assistance with editing the manuscript.

For cross-corpus comparisons a consistent annotation system is necessary. Unfortunately, the currently most widely-used type of grammatical annotation, that of morpheme-for-morpheme glosses, is less suitable for these purposes, for a number of reasons. First of all, morphemic glosses provide no direct or consistent means of identifying syntactic constituents: one cannot reliably and consistently read off the glossing alone, for example, where an NP begins, or where a subordinate clause ends. Nor are there consistently-recognized conventions for identifying "relational" categories, such as subject. Furthermore, morphemic glosses of different languages often use different labels for functionally similar items (e.g. "ACC", or "OBJECT MARKER" for the case marker on direct objects). Thus quantitative comparison of morphemic glosses across different language corpora is exceedingly difficult. Parts-of-speech glossing may provide a basis for cross-corpus investigation, but the research questions that can be addressed are very restricted (cf. Seifart et al. (2010) for discussion). For researchers interested in questions of quantitative, text-based typology, then, there seems to be no other practical alternative than to undertake additional annotation. GRAID is an attempt to provide a cross-linguistically applicable, standardized procedure for such annotations. In order to maximize the possibilities for cross-corpus comparison of GRAID-annotations, it is important to abide by the principles outlined in this manual. The whole point of GRAID is to enable quantitative cross-corpus investigations to be made: this is only possible when annotations in different corpora use the same inventory of symbols, and the same principles for their deployment.

GRAID annotations presuppose considerable finesse in syntactic analysis, and high familiarity with the language concerned. At present GRAID annotations need to be carried out manually which is in principle a time-consuming process. However, for researchers that are familiar with the language concerned, working on texts that they have a thorough understanding of and that often have already been morphologically glossed (see next section), GRAID annotation can be carried out quite rapidly once the annotator has familiarized herself with the system and gained some practice. This is remarkable in view of the fact that generally scholars have been skeptical with regard to the practicability of glossing grammatical relations and clause-level constituents in a language documentation context (cf. Schultze-Berndt 2006).

1.2 GRAID annotations as basis for quantitative analyses

An important feature of GRAID is that the application of glosses yields a count of predicates, arguments and argument positions. The output of a GRAID-annotation is thus a string of symbols that already contains quantitative data that could be used, for example, to answer such questions as:

- What is the ratio of arguments to predicates in a given text?
- How frequent are [+human] expressions in different syntactic functions?
- What is the ratio of covert to overt arguments, what is the ratio of pronominal to NP-arguments?
- Are there regularities of word order of pronominal arguments as compared to NP-arguments?

- Do negated clauses differ from affirmative ones in the way that arguments are realized?
- etc.

Once a text has been glossed, it is a simple matter to extract the GRAID annotation and use any software package that is capable of carrying out complex searches (using regular expressions etc.) on strings of symbols (for example a concordance programme). Preliminary analysis can of course already be undertaken in ELAN, with its somewhat restricted search functions. But the point is that GRAID annotations from different languages, assuming that the annotators have abided by the principles outlined in this manual, provide a basis for direct quantitative comparison of discourse in the languages concerned.

1.3 Prerequisites and design of a GRAID data set

In language documentation projects, software programmes like ELAN or Toolbox are typically used to link annotations of recorded texts directly to the speech signal (time alignment). As a minimum standard in documentation projects, texts are usually transcribed and translated, and often also include a layer of morpheme-by-morpheme glossing, or parts of speech labels (cf. Himmelmann 2006 for an overview of the structure of language documentation projects, and Schultze-Berndt 2006 for annotation practices). This type of pre-annotated text represents the ideal foundation for working with GRAID: a layer of GRAID-annotations can be added to the text, which is intended to complement, rather than replace, the existing layers of annotations.

When working with under-described languages, a minimum prerequisite for a GRAID annotation is an existing transcription, free translation and (at least coarse-grained) morpheme-by-morpheme glossing of the recorded texts and a reasonably comprehensive grammatical description of the language under investigation. This is because GRAID glossing often involves quite subtle, and sometimes quite arbitrary, decisions which need to be **maximally accountable**. Any GRAID data-set therefore requires (a) that the source text with its existing annotation is made available; (b) the annotator formulates an additional short statement in which she makes explicit, and justifies, the analytical decisions made in the GRAID annotation. More specifically, a GRAID dataset should include the following documents for each annotated text:

- ELAN/Toolbox file(s) minimally including distinct tiers for transcription, free translation, morpheme-by-morpheme glossing and GRAID annotations
- sound file (ideally an additional mp3 for ease of access, while retaining the original archived file in a linear, non-compressed format)
- text document containing only transcription and free translation (preferably pdf)
- text document containing transcription, morpheme-by-morpheme glossing, GRAID annotations and free translation with morpheme and GRAID glosses being left-aligned common with morphemic glossing (pdf)
- text document containing export of GRAID glosses (plain text)

1.4 Changes since Version 6.0

The main practical change concerns the implementation of GRAID annotations within other layers of annotation: while these were previously only loosely associated with entire utterance or clause units, we now enter glosses in a word-for-word manner, hence arriving at a more systematic alignment of GRAID glosses with particular constituents. This change has already been outlined in a working paper ('Annotation with GRAID in ELAN: draft guidelines for RP-project', circulated by Haig & Schnell in October 2012), but some of these guidelines have become obselete as well.

Otherwise, the basic features of the GRAID system have proven practicable, and have not been changed significantly. The main other changes include:

- Adoption of a more articulated annotation of different clause types (cf. 2.6).
- Lower priority afforded to the argument/adjunct distinction than in previous versions. In practice, this mostly affects the annotation of the function glosses locationals <l>, goals <g>, and <obl> (cf. Section 2.4 for exemplification and discussion).
- Introduction of the possibility of annotating phrasal sub-constituents (i.e. identifying non-head elements, and allowing them to be assigned to a particular phrase; cf. Section 2.6).
- Adoption of a hierarchical approach, allowing for differing levels of details in certain areas, depending on the interests of the researcher and the peculiarities of the language, while still maintaining cross-corpus comparability.
- Greater attention to the annotation of bound person markers (cf. Section 3). Thus we now explicitly endorse the possibility of annotating two distinct exponents of the same argument within a particular clause, under certain conditions. This remains perhaps the most contentious issue, and our modifications in this regard have also been prompted by recent developments in the literature (cf. Haspelmath 2013).
- The addition / modification of certain symbols or symbol combinations. As outlined in previous versions, we attempt to keep the symbol inventory to a minimum, in keeping with general demands for generality and simplicity. There is an obvious temptation to create new symbols as a response to each language-specific problem, which we have resisted wherever possible; nevertheless, some additions seemed, in view of their cross-linguistic frequency and saliency, justified.
- Furthermore, we briefly outline possibilities for combining GRAID annotations with referent indices, in order to facilitate investigations of accessibility and persistence in reference tracking (cf. Ariel 1990, Chafe 1976; 1994).

2 Using GRAID annotations in practice

In this section, we first provide a short outline of GRAID in order to ptovide the user with a feel for the basic workings of the system, before giving explications

of the full inventory of symbols and more extensive examples. In Section 3, we discuss issues of linguistic analysis and annotation connected to argument indexing and so-called 'agreement'. Section 4 then deals with other specific issues of analysis and provides practical guidelines for annotators.

2.1 Overview

Throughout this manual we enclose the actual symbols of GRAID in triangular brackets, like this: <...>. GRAID glosses are, by convention, aligned with single words, but target entire referential expressions and their functions (e.g. argument functions S, A, P, adjuncts, etc.), i.e. phrases rather than words. Essentially, each item of a GRAID annotation couples an abbreviation for a form (e.g. <pro> 'full pronoun'), which may additionally have an animacy feature, e.g. 'human', with a function, e.g. <s>. Animacy features such as 'human', which semantically specify individual form units, are linked to forms with a full stop, while forms are linked to their functions via colons <:>. An example is the first constituent of (1):

(1) ***he*** *is* *leaving now*
pro.h:s aux v:pred other
(= full pronoun, human referent, in S function)

As indicated above, GRAID annotations involve decisions on what elements are part of the same prosodic word, and whether they have affixal or clitic status (cf. the discussion on wordhood in e.g. Dixon and Aikhenvald (2002), and the research by Bickel and associates on the typology of the word (cf. Schiering et al. 2010)). We follow the conventions of the *Leipzig Glossing Rules* (Comrie et al. 2008) for distinguishing between clitics and affixes: affixes are linked by a dash <->, while clitics are linked by <=>, as illustrated in the following examples (cf. Section 2.2 for more details on formal properties of clause constituents):

(2) a. Yagua, Lowland Peru, unclassified, Payne (1992:18)
Sa*-jutu-****rà***
3s.A-carry-**3sg.inan**
pro.h:a-v:pred-**pro:p**[2]
's/he carries it'

b. German
er ***hat's*** *gemacht*
he **has=it** do:ptcpl
pro.h:a **aux=pro:p** v:pred
'he has done it'

As can be seen from the examples above, the glossing of predicates is rather coarse-grained, since the main target of GRAID annotations is the realisation of referential expressions. The main verbs are simply glossed <v:pred> 'verbal predicate', likewise consisting of a form and a function tag. Different types of predicate and glossing conventions are outlined in Section 2.5 below.

[2] Prosodic word consisting of: pronominal affix with human referent in A-function + verbal predicate + pronominal affix, non-human referent, with P-function.

In many languages, in particular those classified as morphosyntactically 'isolating', referential expressions typically consist of more than one word. Isolating Oceanic languages represent this type of language: these often have articles and TAM markers that obligatorily introduce NPs and verb complexes, respectively. In such cases, the GRAID gloss for the entire phrase will be aligned with the lexical head word of the phrase. An example from Vera'a illustrates this:

(3) *[...]* ***ne*** ***kal*** ***'ō'*** ***ba'a*** ***kēl*** ***sar*** *ēn* *'añsara* *ē*
TAM2:3SG enter carry inside back in ART person DEM3
\# 0.h:a v:pred np.h:p
lē *=n* *mē'ērsa*
LOC =ART harbour
adp np:g
'... and then he took that man back ashore at a harbour.' ISAM.065

In this example, the verb complex is introduced by a TAM particle, followed by the head verb which is in turn followed by further verbs and adverbs. Only the head verb is noted in the GRAID glossing and receives the function gloss for the entire phrase. Similarly, the object NP is introduced by an article, followed by the head noun and a demonstrative; only the head noun receives GRAID glossing. The other words within each of these two phrases can be left unglossed as this example, or optionally be glossed as sub-constituents. This practice is outlined in Section 2.8 below. As can also be seen in this example, adpositions are noted separately, while the head of the complement NP carries the function gloss for the entire PP.

A further point to be observed in this example is that unexpressed arguments are noted in GRAID annotations (the AGENT argument human reference of 'carry' is glossed as <0.h:a>). The glossing of unexpressed arguments is of course a delicate matter, and crucially, it presupposes that the annotator has a sound notion of which type of argument belongs to a particular predicate. The challenge then is firstly to determine whether a particular predicate expression (verb, series of verbs, or other) in fact licenses the non-expressed argument, and secondly, whether the argument position refers to a discourse-retrievable entity. Only if the answer to both questions is 'yes' do we recommend to gloss a zero argument. Their form is rendered in the gloss by <0> (the digit 'zero'). In a configurational language like Vera'a, the unexpressed argument is fairly straightforwardly associated with a particukar slot in linear order, and hence the GRAID gloss is aligned with this 'empty slot'. When using ELAN, it is often desirable to insert an empty annotation in the object language tier at the positin where the zero argument can be assumed.

In languages with freer word order, however, it is obviously not particularly meaningful to assume a particular linear position of a 'zero element'. We generally insert zeroes in the position where the argument concerned would, in a pragmatically neutral clause, occur. This is not an important issue, and in principle, investigators could decide on some arbitrary standard position within the clause annotation for positioning zero arguments. The important point is simply that they need to be represented within the boundaries of the clause at some position. As illustration, consider the following sequence from Northern Kurdish, clauses (4a) and (4b) have one zero argument each, while clause (4c) has two:

(4) Northern Kurdish

a. *t-ere* *cî-k-î* *dûr,*
INDIC-go.PRS-3SG place-INDEF-LNK distant
0.h:sv:pred np:g other
'(he) goes to a distant place'

b. *0* *ew-ê* *sandiq-ê* *jî* *di-gr-e*
dem-OBL box-OBL too INDIC-take.prs-3sg
0.h:a other np:p other v:pred
'(he) takes that box as well'

c. *0* *0* *d-avêj-e* *behr-ê*
INDIC-throw.PRS-3SG sea-OBL
0.h:a 0:p v:pred np:g
'(he) throws (it) into the sea'

A further issue with zero arguments is that they must generally be substitutable by an overt form. Thus in (3) and (4) above, the subject could have been realised as a NP or free pronoun without any impairment of grammaticality. Certain types of predicate, though, imply a referential argument, but the overt expression of that argument is systematically suppressed. This is the case with various types of non-finite predicate, which head clause-like phrases, but do not permit, for example, the overt expression of S or A within the clause. In such cases, we follow Bickel (2003) in not glossing the unexpressed argument with <0>, because speakers have no choice at this point. A special gloss is provided for such predicates, <vother>, which is discussed in Section 2.5.5 below.

The basic unit for glossing is a clause, defined here as the entirety of constituents associated with a particular predicate. Obviously defining clause boundaries is not always straightforward; some problems are discussed below in Section 4.1 in connection with the counting of predicates. GRAID signals the left-handed boundary of syntactically independent, main clauses with <##>; to this, various modifiers can be added. Syntactically dependent clauses are marked with <#>, again to which modfiers are added (cf. 2.6 below for details).

In the following sections we provide the complete inventory of GRAID symbols and explain their uses. Symbols are divided into three main categories: symbols indicating the forms and inherent properties of referential expressions, symbols for their functions, and symbols for glossing predicates. Finally, we introduce some additional symbols for certain clause types and uncertain cases. In Section 5, a full alphabetical list of all symbols used may be found.

2.2 Formal properties of referential expressions

The core of GRAID annotations is the glossing of referential expressions. Moreover, GRAID annotations focus on the glossing of verbal arguments, which means that not all word/form classes typically attested across languages receive detailed treatment (see Section 2.4). The main symbols used for the form of referential expressions are contained in Table 1.

The distinction betwen NP, pronoun and zero is the most central one for GRAID annotations. A few remarks on how to employ these are necessary: the

Table 1: Glosses for the form of referential expressions

np	noun phrase
pro	free pronoun in full form
=pro	'weak' clitic pronoun
-pro	pronominal affix, cf. 3
0	covert argument / phonologically null argument
refl	reflexive or reciprocal pronoun, cf. Section 4.2
adp	adposition
w	'weak' (optional symbol), indicates a phonologically lighter form, it precedes the form symbol, e.g. <wpro>
x	'non-referential', see below for explanation
other	used for expressions 1. that are not of a type listed above 2. the form of which is not considered relevant
ln	NP-internal subconstituent occurring to the left of NP head
rn	NP-internal subconstituent occurring to the right of NP head
lv	subconstituent of verb complex occurring to the left of verbal head
rv	subconstituent of verb complex occurring to the right of verbal head

label <np> is basically intended to capture what in the literature is often labelled 'lexical mention/expression/etc.' (cf. Du Bois 1987 among many others), and these wil typically refer to those NPs headed by common nouns. In practice, however, this category will also include expressions like personal or place names and phrases, certain kin or address terms, numeral expressions ('one of them'), etc. Obviously, many of these types of expressions evoke a plethora of theoretical issues, which we won't touch upon here. Where annotators wish to preserve distinctions like the ones just mentioned, these should be added in a way to be outlined in Section 2.9 below, so as to not impair the original plain GRAID glossing.

Similarly the category labelled <pro> bears obvious complications. First of all, note that this label is intended to capture the core of what Lyons (1968:268) calls 'definite' pronouns, i.e. forms typically expressing categories like person, number, gender, clusivity, honorificity and having given-activated reference. This then excludes indefinite, interrogative, etc. pronouns. These latter elements are not captured in our system, and some of them may in fact be treated as NPs in some cases. Again, annotators may in principle preserve relevant distinctions by introducing language-specific tags not included in the core inventory described here. Secondly, what may be glossed as 'pronoun' may come in various forms. Of particular concern here are various forms of what is sometimes called 'bound pronouns', and 'clitic pronouns'. It was already said above that we basically exclude canonical cases of person agreement from glossing, and only recommend using the <pro> gloss for those bound person markers that Haspelmath (2013) calls 'cross-index' and 'pro-index'. We will explain below how degrees of boundedness are noted in GRAID. A more detailed discussion on the treatment of bound person markers and their co-occurrence with free pronouns and NPs is presented in Section 3.

The symbol <0> is probably the most controversial. The 'existence' of zero argument as such is a quite delicate issue, and although most fellow linguists would probably assume that the category is valid and useful, many will hesitate

to make definite decisions in specific instances. As guidelines we propose three conditions for the assumption of a zero argument. First, annotators need to decide for a given clausal construction which arguments are required by the predicate. We assume that annotators familiar with the language concerned will basically be in a position to make such decisions regarding the valency of predicates, e.g. by checking for a given lexical verb in the lexical database for examples of attested argument structures with overt expressions.

The second condition is that the argument in question be expressable by an overt form in a given construction, so that a pronoun or NP could occur instead of zero without violating grammaticality. Thus, in cases where an argument role is systematically suppressed we recommend not to gloss <0> because there is no alternative to this 'form'. Typical examples are the suppression of e.g. S or A argument in non-finite clause constructions, like participle or converb constructions in many languages. Examples of zero arguments are cases in narrative texts where the reference of arguments is considered inferrable from the discourse context, as in examples (3) and (4) above. Equi-deletion in English also falls under this category: in the sentence *Peter works in London but (he) lives in Cambridge*, the pronoun *he* could be omitted, in which case we would gloss <0.h:s> 'zero, standing for a third person argument with human referent, in the S-function'. The third condition concerns the referentiality of the omitted argument. In a sentence such as *We'll find a restaurant and eat there*, we would not gloss a <0:p> for an "omitted" object of the verb *eat*, because in this context it refers to an activity with inherently understood, but unspecified, object. If no clear reference for the omitted argument is available from the context, then we consider it non-referential, hence do not gloss it. Another example would be the instrument role of a predicate meaning 'cut', where we would not assume a zero argument on the rather dubious assumption that the action of cutting *necessarily* involves the use of some appropriate instrument (scissors, knife, ...). Thus, what is relevant is whether a specific entity can be said to be involved in a state of affairs without being overtly expressed in the respective clause construction, not whether a particular role is generally evoked by the semantics of a lexical item. In practice, we have found the guidelines to suffice for most contexts. Where annotators are unable to reach a clear decision, they may gloss the entire clause unit with <nc> 'non-classifiable', which would exclude that particular clause unit from the analysis. This is an option generally available for cases of uncertainty (see below).

As mentioned above, the hyphen <-> and equal sign <=> indicate affixal and clitic boundary respectively. These are most commonly used in GRAID with bound person markers; however, they may also be optionally used with canonical agreement morphology (cf. Section 3) or with incorporated nouns in polysynthetic languages (cf. (16) below). As for distinguishing clitics from affixes, we follow Bickel and Nichols (2007) in assuming subcategorization to be the primary diagnostic: if the elements concerned are restricted to hosts of certain classes, they are affixes, if they are not, then they are clitics.

Languages may have three grades of phonological weight for certain forms, for instance pronouns (like French *moi, je, j=*), and researchers must make a decision on which of the three are to be considered free and which are to be considered clitic. An option that GRAID allows for is the additional letter <w> 'weak' that can be added to <pro> yielding <wpro> if annotators wish to preserve a three-way distinction (another typical candidate is <aux> 'auxiliary',

yielding <waux>; cf. Section 2.5.2 below).

Where arguments are marked by a preposition or postposition, these are glossed with <adp>. The glossing of the function of the **entire adpositional phrase** is noted, however, on the NP. An example is the following, where the function gloss <l> refers to 'locative' (see next section):

(5) German

Sie	*wohn-t*	***in***	***diesem***	***Haus***
she	live-PRES.3S	in	DEM.NEUT.DAT	house
pro.h:s	v:pred	**adp**	ln	**np:l**

'She lives in this house'

The gloss <other> is used for any expression that is neither a NP nor a pronoun. Typical candidates for glossing with <other> are adverbs or particles. In such cases, it may be combined with a function gloss so that, for instance <other:g> may be used for a locative adverb denoting the goal in a motion event:

(6)

they	*ran*	***uphill***
pro.h:S	v:pred	**other:g**

The symbol <x> is not strictly speaking a form gloss. Rather, it marks those forms that appear in a construction for idiosyncratic reasons and do not refer to anything. A typical case are expletive subject pronouns in some Germanic languages (Engl. *it*. German *es*), glossed <xpro>. Some instances of NPs may also be thus marked, e.g. the object NP in a idiomatic construction like *He kicked the bucket*: this NP does not have a specific referent and thus—to be consistent with our assumptions about zero arguments—needs to be marked off in some way[3], yielding <xnp>.

The uses of <refl> are discussed in Section 4.2 below.

Finally, the symbols <ln>, <rn>, <lv> and <rv> are optionally used where annotators wish to take note of phrase-internal subconstituents in addition to clause-level constituents. The glosses <ln> and <ln> are used for subconstituents preceding the functional head of the phrase (i.e. occurring to its left), and <rn> and <rv> for elements occurring to its right. These form indexes do not combine with function glosses, but merely take a symbol that indicates the type of phrase they occur in, e.g. <lv> would be an element occurring in a verb complex, preceding the verbal head, as is the case with TAM markers in Vera'a, cf. (3). The details of subconstituent glossing will be discussed in Section 2.8 below.

2.3 Semantic and pragmatic properties of referential expressions

We now turn to the symbols for semantic and pragmatic properties of arguments. An overview is given in Table 2.

The properties discussed here basically comprise person, and animacy (human vs. non-human). These are linked to the form glosses using the symbol

[3] We thank Ruth Singer for directing our attention to such cases of non-referential object NPs.

Table 2: Glosses for the properties of referents

1	1st person referent(s)
2	2nd person referent(s)
h	human referent(s)
d	anthropomorphized referent(s); the use of this symbol is optional

<.>, as demonstrated in (1) above. The bare <np> or <pro> symbol is used where a NP or pronoun has a a 3rd person non-human referent. With 3rd person human referents, <np.h>, <pro.h> and <0.h> are used. For 1st and 2nd person referents, <1> and <2> are used respectively. As humanness is entailed in reference to speech act participants <h> would be redundant in combination with <1> and <2>, and is therefore not used.

In some languages, pronominal forms with a particular paradigmatic person value may be used for referents of a different person in some contexts. In cases of e.g. German *Sie*, we suggest going with actual reference to an addressee and gloss <pro.2> for 2nd person.

The symbol <d> is optionally used with anthropomorphised discourse participants (e.g. <np.d>). It is intended to distinguish e.g. deities, spirits, mythical figures, capable of speech and self reference, from genuine human discourse participants, if the researcher believes the distinction may be syntactically or otherwise linguistically relevant. Our experience until now has been that the distinction human vs. non-human is the most relevant one in accounting for attested variation. Nevertheless, for some languages researchers may find it necessary to make finer-grained distinctions, and additional symbols could then be used in this slot. Again, additions should be made with caution and only where absolutely necessary, and should not impair original GRAID glosses. The exact use of additional symbols should be noted in the documentation.

2.4 Syntactic functions of referential expressions

GRAID annotations link symbols for forms, as introduced in the preceding section, with symbols for syntactic function, using the general format <form.animacy:function>. In this section, we summarize and exemplify the symbols for syntactic functions, a term we use largely interchangeably with the term 'grammatical relation' here. We focus on the major syntactic functions S, A and P (in actual glosses we also use the small case letters, whereas in the text discussion we use the upper-case letters; this minor inconsistency can be ignored).[4] Additional function labels include <poss> 'Possessor' and <g> 'Goal', which are discussed below. Syntactic functions are intermediate between language-specific cases (nominative, accusative, genitive etc.) and thematic roles such as Agent, Experiencer or Theme. Syntactic functions enter different grammatical relations defined via their morpho-syntactic behavior (Bickel 2011; Andrews 2007). Although the precise theoretical status of syntactic functions and grammatical relations remains controversial, a considerable body of research suggests that they do represent a valid level of syntactic description and, more importantly, provide a framework within which significant

[4] Note that Andrews (2007:139) distinguishes the grammatical *functions* S, A and P from grammatical *relations*, e.g. subj and obj. The former generally subsumes the functions S and A on grounds of common marking and/or behavioral properties.

cross-linguistic generalizations on the possible shapes of grammars can be formulated (Comrie 1989, Farrell 2005, Andrews (2007), Haspelmath 2011 among many others). Crucially, this level of syntactic organization is generally neglected in most conventional glossing procedures. Table 3 gives an overview of the functions recognized in GRAID.

Table 3: Glosses for major syntactic functions

s (or: S)	intransitive subject
a (or: A)	transitive subject
p (or: P)	transitive object
ncs	non-canonical subject
g	goal argument of a goal-oriented verb of motion, but also: recipient of verb of transfer, and addressee of verb of speech
l	locative argument of verbs of location
obl	oblique argument, excluding goals and locatives
p2	secondary object
dt	dislocated topic (right or left-dislocated)
voc	vocative
poss	possessor
appos	appositional
other	other function

The symbols for syntactic functions combine with the symbol(s) for form and semantic properties in Tables 1 and 2 to yield composite labels. Typical examples of frequent combinations are, e.g.:

<pro.1:a>	'first person pronoun, in A-function'
<np:l>	'lexical noun phrase indicating location'
<=pro.2:poss>	'clitic pronoun, second person, indicating the possessor'
<0.1:g>	'unexpressed first person argument, recipient or addressee'
<np.h:poss>	'full NP with human referent, possessor function'

Further examples are shown in context in (1)–(6) above. For identifying S, A and P, we essentially follow the approach of Andrews (2007:137f.): A and P are those arguments of a transitive verb that receive the same formal coding as Agent and Patient of a primary transitive verb denoting a prototypical transitive event (e.g. English *kill*, *smash*) in the language concerned. S is used for the sole arguments of intransitive verbs, including the subjects of non-verbal or copular clauses. Crucially under this view, only those clauses count as 'transitive' that have both an A and a P argument. Consequently, where an agent-like argument co-occurs e.g. with an oblique argument in a two-argument clauses, this argument will be glossed S rather than A. This contrasts with recent views such as Dixon (2010:151), who extends A and O (=P) to arguments not marked in the same way that the A and O (=P) of primary transitive verbs are. We nevertheless prefer the restricted view, according to which A and P are reserved for those arguments coded identically to the core arguments of primary transitive verbs. Put differently then, syntactic functions as perceived here neutralise particularities of semantic roles on formal grounds rather than semantic abstractions; they are grammatical functions rather than semantic macro-roles (cf. Bickel 2011, Van Valin and LaPolla 1997 for semantic macro-

role approaches to similar categories; and Haspelmath 2011 for discussion)[5].

Similarly, the S role is often understood in a broader sense. Typically, the S argument takes the form of either A (accusative alignment) or P (ergative alignment). However, S is sometimes also used for the single argument of any monovalent verb, regardless of its overt form, (e.g.Donohue (2008) refers to the dative EXPERIENCER of a verb of physical perception as S; cf. discussion below). We are unaware of any attempt to define a suitable 'anchor' for identifying S in a given language; most scholars simply take S to be the "single argument of a one-place predicate". We suggest that for identifying <S>, the form of subjects of declarative, affirmative, present-tense statements involving simple property-assignment predicates should be taken as a benchmark, e.g. 'be big', or 'be black' (excluding, of course, expressions of physical sensations). For the vast majority of languages known to us, subjects of this kind of predicate will be in the formally least-marked form available in the language (e.g. a nominative or absolutive case, if available).

For arguments marked differently from S, A or P in the language, GRAID offers varying options. One quite common argument type are those which evidently share syntactic properties of S and A, but differ in their case marking. For such arguments, we suggest the gloss <ncs> 'non-canonical subject'. The dative subject in the following Icelandic sentence could be glossed as follows:

(7) ICELANDIC

mér	*er*	*kalt*
1SG.DAT	is	cold
# **pro.1:ncs**	cop	other:pred

'I feel cold'

In addition to these cases of core argument, we recognise three types of oblique arguments. The first are locatives, <l>. This symbol is used for oblique arguments expressing local roles of static location, and also source. A typical example of such a locative argument was presented in (5) above, repeated below:

(5') GERMAN

Sie	*wohn-t*	***in***	***diesem***	***Haus***
she	live-PRES.3S	in	DEM.NEUT.DAT	house
pro.h:s	v:pred	**adp**	ln	**np:l**

'She lives in this house'

For local 'goals', entailing a change in position or movement in a specific direction, we use <g>. Of course languages frequently extend the formal means for indicating local goals to RECIPIENTS and ADDRESSEES; in such languages, all three will be glossed <g>, this gloss thus covering functions extending beyond the semantic role label GOAL. Note that in these cases, RECIPIENTS and ADDRESSEES would receive a gloss for animacy, e.g. <pro.h:g>, so that the

[5] Our conception of S, A and P is considerably more restricted than that of Bickel and associates (e.g. Bickel and Nichols 2009), which draws on a proto-role-based approach. The differences across different concepts of S, A and P have recently been critically summarized in Haspelmath (2011), who also proposes a semantic "anchor" type for the S-role. We refer readers to that paper for the details of the different approaches; here we simply note that the usage of these terms is far from uniform in the literature, hence the need for explicit definitions.

distinction between them and purely local goals would still be be recoverable. Other languages, however, systematically distinguish the expression of GOALs from that of RECIPIENTs, the latter typically also encoding ADDRESSEEs. In such cases, there are two options: the <g> gloss could be used for RECIPIENTs and ADDRESSEEs too, which would obviously gloss over some language-specific details. Alternatively, the <g> gloss could be reserved for local GOAL arguments, while the others would be glossed <obl> (see below).

In some languages, the locative roles GOAL and LOCATION may both be encoded in the same way by means of a general locative case marker or adposition. Other languages formally distinguish between GOAL and LOCATION. In the former case, again, annotators have to decide whether they gloss both GOALs and LOCATIONs with <l>, that is, taking language-specific marking properties at face value, or whether they consider it more important to capture the semantic difference between the two roles.[6]

The <obl> gloss is used for arguments that cannot be subsumed under S, A, P, <g> or <l>. A typical case of an oblique argument for which glossing as <g> or <l> is certainly not an option is the following where the prepositional phrase expresses a theme and the direct object the recipient argument:

(8) ENGLISH (Andrews 2007:158)
They provide us with weapons.
pro.h:a v:pred pro.1:p **adp np:obl**

In this example, and in similar cases, the direct object is still glossed as <p>, due to its formal identity with other direct objects expressing typical PATIENTs, and despite its semantic role as RECIPIENT. Other examples for <obl> are the dative complements of German *helfen* 'help', or the instrumental-marked complement of Russian *vladet'* 'master, rule' (the 'Exceptional Case Marking' of earlier versions of Generative Grammar). In GRAID, these NPs would receive the function-gloss <obl>. Further examples are verbs expresssing concepts such as 'meet', which may require a COMITATIVE complement coded in a manner distinct from a P. Essentially then, <obl> is the gloss of choice for non-term arguments, i.e. those that differ formally from <p>, and are not <l> or <g>.

This is an area of considerable complexity, and annotators need to decide early on which solution they wish to adopt, and apply it consistently. However, our experience has shown that the three categories <g>, <l> and <obl> do in fact provide the basis for a working solution for the glossing of non-core arguments.

The problems with distinguishing arguments from adjuncts have occupied linguists for decades. However, for the practical purposes of glossing spoken narrative texts, we have found the theoretical discussion to be of surprisingly little relevance. The point can be illustrated by examples of locational arguments. Consider the following German examples:

[6] Note that in the case of GOALs versus RECIPIENTs/ADDRESSEEs, coding the latter as <obl> due to distinct marking properties in turn results in including these arguments with other oblique arguments from which they differ semantically, and also formally. For example, the THEME argument of English *supply* can be flagged by the preposition *with*, as in *They supplied us with weapons.* Hence, annotators have to make a decision here and document it in the notes.

(9) German
   a. Sie liegt auf dem Sofa.
   b. Sie wohnt in München.
   c. Sie wohnt möbliert / ruhig / zur Miete.
   d. Sie arbeitet in dem Büro.

For (a), it has been a matter of some debate whether locational expressions are (obligatory or optional) arguments of postural verbs, or merely adjuncts, and whether a semantic shift on the verb is triggered by its omission, and so on. For the verb 'live/dwell' in (b), it has been suggested that the locational complement should be considered an argument of the verb *wohnen*, because it cannot be ommitted here, cf. the ungrammaticality of **sie wohnt* 'she dwells'. However, (c) shows that the complement need not be locational; thus, the verb *wohnen* seems to require some complement, and a broad range of forms and expressions is acceptable. For (d), on the other hand, it could be maintained that the PP expresses a purely circumstantial expression, merely adding information on location to the clause, but not required by the semantics of the verb. But in practice, particularly when dealing with under-described languages, all these decisions turn out to be problematic (if not plain arbitrary) for a number of reasons. And notably, it does not seem to be the case that languages specifically distinguish 'argument' locatives from 'adjunct' locatives in their morphosyntax; thus in German, all kinds of locationals are expressed through prepositions of one sort or another, with no obvious correlation with argument vs. adjunct. The conclusion that we draw from these, and similar, issues is that it is better practice to treat all expressions of static location as <l>, regardless of putative argument/adjunct distinction. Once an annotation is completed, it would then be possible to re-examine all the <l> glosses to determine whether specific valency patterns can in fact be detected. But initially, we recommed glossing all locationals and goals with <l> or <g> respectively.

A somewhat different set of problems arises in the case of primary-object constructions, where Recipients are coded as P, cf. Malchukov et al. (2010). An example from English is the following:

(10) English
   Mum gave ***us*** *sweets.*
   # np.h:a v:pred pro.1:p np:p2

Here the pronoun *us* is a Recipient, but it is coded in exactly the same way as the Patient argument of a primary transitive verb in English, and is thus analysed as a P argument. In this case, as in general in GRAID, **formal morphosyntactic coding properties take precedence over semantics**, and the Recipient will be glossed as <p>. The secondary object *sweets* expressing the theme would receive the gloss <p2>, thus representing a subclass of P arguments. Note that when analysing GRAID glosses, the two types of P arguments can be collapsed.

Certain problems also arise with the glossing of syntactically 'ambiguous' elements, such as *me* in the following example:

(11) English
   He expected ***me*** *to leave.*

This example, and related issues, are taken up in Section 4.4 below.

Expressions for circumstantials may be simply glossed as <other> without any specifications of form and function. A further possibility is to analyse its form, but give its function as <other>, yielding for instance <adp np:other>:

(12) English

In winter *she lives in that house*
\# **adp np:other** pro.h:s v:pred adp ln np:l

The symbol <dt> 'dislocated topic' is used for NPs that are either fronted to the clause proper or occur at the right clause boundary, and do not have an argument relation in the clause. Colloquial English makes extensive use of such elements, as in (13):

(13) English

Mike, *he hates syntax.*
\# **np.h:dt** pro.h:a v:pred np:p

Note that we use the term 'topic' here because one of the functions of the classical cases of *left*-dislocation is often used for 'topic announcing' or 'frame setting' functions (Lambrecht 1994; Chafe 1976 for discussion of these two functions). However, the use of this term is somewhat lax, as it will also be applied to *right*-dislocated 'anti-topics' (cf. Lambrecht 1994) and 'topicalised' phrases that are otherwise pragmatically marked. A further option for the glossing of dislocated phrases is to also note the clause-internal function the phrase correlates with:

(13') English

Mike, *he hates syntax.*
\# **np.h:dt_a** pro.h:a v:pred np:p

Hence, the function gloss <dt_a> here stands for 'dislocated topic(alised) corresponding to A function', the underscore separating the standard <dt> from the additional gloss for A function. Where no additional function is noted, the phrase does not have a corresponding argument within the clause. Annotators should note in their documentation whether they do note additional clause-internal functions with dislocated phrases, as this will obviously alter the meaning of the bare <dt> gloss. The consideration of further categories and distinctions will be discussed in more detail in Section 2.9.

A further gloss for a non-argument function is the <voc> 'vocative' gloss. The term is obviously adopted from the vocative case as attested in Classical Latin or South Slavic languages and is used for referential expressions used deictically in order to evoke the attention of a particular person.

(14) English

Tim, can you get the door?
\# np.h:voc aux pro.2:a v:pred ln np:p

The possessor function of pronouns, nouns or NPs can be glossed as <poss>. Possessors may express semantic roles like Beneficiary or Recipient in some languages, for instance, in Oceanic languages (cf. Margetts 2004, 2007). But even where possessors are embedded in a possessive NP, they may have an impact on information packaging and discourse structure, and provide the anchors

for anaphoric reference or control constructions (cf. for instance ***my*** *plan was to leave the party early and go swimming with Emily*, where the possessive pronoun *my* provides the reference for the unexpressed subjects of *leave* and *go swimming*).

The gloss <appos> 'appositional' is used for expressions in apposition to another one. The two expression bear the same syntactic function; but in contrast to dislocated expressions, two appositional expressions would occur *within* a single clause.

(15) English

	Tim,	*the butler,*	*opened the door.*		
#	np.h:a	ln	np.h:appos	v:pred	ln	np:p

For other functions that do not match those mentioned so far there is the option of glossing them with <other>.

2.5 Predicates

In accordance with the research questions outlined in Haig and Schnell (2009) (cf. also Haig et al. 2011a; Schnell and Haig, submitted), the glossing of predicates is less elaborate. Form and function symbols used specifically for the glossing of predicative expressions are given in Table 4. Other glosses already introduced above are also used with predicates, as will become clear in the following sections.

Table 4: Form and function glosses for predicates

v	verb or verb complex (cf. Section 2.5.1)
vother	non-canonical verb-form (cf. Section 2.5.5)
cop	(overt) copular verb (cf. Section 2.5.2)
aux	auxiliary (cf. Section 2.5.2)
-aux	suffixal auxiliary
=aux	clitic auxiliary
pred	predicative function
predex	predicative function in existential / presentational constructions

A broad distinction is drawn between clauses containing a verbal, copular or non-verbal predicate. We will briefly discuss each type in the following sections.

2.5.1 Verbal predicates

Where a phrase functioning as the predicate is classified as 'verbal', it will be glossed with <v:pred>, as in most examples considered above. Verbal phrases (cf. for example the 'verb complex' in Oceanic) functioning as verbal predicates typically have a lexical verb as their head and show TAM morphology either on the head word itself or elsewhere (e.g. TAM particles, as in Oceanic). For languages that apparently lack verbs as a distinct lexical class (e.g. Kharia (South Munda), Peterson 2011), annotators will have to decide whether a particular predicative phrase can nevertheless be analysed as 'verbal' (e.g. in Kharia due to TAM marking occurring as enclitics, cf. Peterson 2011), and may thus use the gloss <v:pred> as the default gloss for such predicates. Alternatively, they

may prefer to use <other:pred>. The use of the <vother> gloss is discussed in Section 2.5.5 below.

As with glosses for referential expressions, the <v:pred> and other possible glosses are associated with the entire phrase serving as the predicate. However, in the word-for-word glossing format, these glosses will align with the (lexical) head word of the predicative phrase, thus typically a lexical verb. Other elements occurring within the predicative phrase, e.g. TAM-marking particles or valency-changing devices, will be glossed as subconstituents, cf. Section 2.8 for exemplification. Further elements thus covered are serialised verbs, light verb constructions, auxiliary verbs or certain types of adverbs. Here investigators must reach language-specific decisions on whether to treat additional elements as part of the same predicate (hence receiving a subconstituent gloss), or a distinct element on clause level, receiving some other gloss. There are usually morphological, syntactic and semantic arguments in favour of one analysis over the other, which should be made explicit in the additional documentation. Ideally, these issues will have been investigated in some detail in the grammatical description of the language, to which annotators should make reference.

Where predicative phrases contain referential information, as in the case of (more or less bound) person markers (cf. Section 3 for examples), or incorporated nouns, they will either be linked with <-> or <=>, or be glossed as a subconstituent of the predicative phrase, depending on their morphological status. Exs. (16a) and (16b) contrast a transitive sentence with a free NP in P function and the corresponding clause with incorporated P:

(16) Huahtla (Nahuatl, Uto-Aztecan, Mexico) (Mithun 1984:860)

a. *neʻ ki-caʻ-ki kallak-tli*
he it-close-PST door-ABS
\# pro.h:A v:pred np:P
'He closed the door'

b. *neʻ kal-caʻ-ki*
he door-close-PST
\# pro.h:A np:P-v:pred
'He closed the door'

(17) Vera'a

a. *dir =k ʼēn sar ma =n sava […]*
3PL =TAM2 see in hither =ART what
\# pro.h:a =lv v:pred rv rv =ln int_np:p
'Then they spotted something inland.' JJQ.256

b. *dir =ēk qērē ba'a di sar lē =n*
3PL =TAM2 push into 3SG in LOC =ART
\# pro.h:a =lv v:pred rv r_pro.h:p rv adp =ln
mō-gi =n nimnē
POSS.DVEL-3SG =ART house
ln =ln np:g
'They would then push her into her house.' ISWM.171

2.5.2 Copular predicates and auxiliaries

'Copular' verbs are members of the class of verbs, but they are largely devoid of lexical semantics (and are often defective and/or highly irregular). Copulas serve to carry inflectional morphology, and functionally link a subject NP to some other kind of phrase (e.g. AP or other NP). The relationship between the two phrases is generally one of four kinds:

(18) a. Identification/Equation:

That woman is the managing director
ln np.h:s cop ln ln np.h:pred

b. Classification:

Bob is a linguist
np.h:s cop ln np.h:pred

c. Property assignment:

He is very old
pro.h:s cop other:pred

d. Location:

She is at the market
pro.h:s cop adp ln np:l:pred

In English, a form of the copular verb *be* is used for all four types. We gloss an overt copula, such as *is* in the English examples, with <cop>. The element which is the semantic predicate, on the other hand, recieves the function gloss <pred>, and whichever form gloss is appropriate. For example, in (18b) we could gloss <np.h:pred>. If none of the available form labels are suitable, the predicate can be glossed <other:pred>.

The class of copular verbs is not always clearly demarcated. With doubtful examples, such as *become* and its equivalents in other languages, annotators may choose to gloss the entire predicative phrase like a regular verbal phrase, i.e. <v:pred>, yielding:

(19) *She got hurt*
pro.h:s v:pred rv

Languages often have predicative constructions that involve a copula verb and bear existential or presentational reading. We discuss different types of construction with existential or presentational function in Section 2.5.4 below.

We consider auxiliary verbs to be verbs bearing information on tense, aspect and mood, but which do not impact on argument structure (i.e. do not license grammatical relations). They are glossed <aux> where occurring outside of the verbal phrase, and <l_aux> or <r_aux>, i.e. as subconstituents occurring either to the left or the right of the head within the verbal phrase.

The formal properties of copulas and auxiliaries can be marked in the same way as with pronouns, using <->, <=> or <w...>.

2.5.3 Non-verbal predicates

In many languages, predicative expressions such as those illustrated in (18a)–(18d) above, do not require any overt verbal element. Instead, NPs, APs or PPs expressing the predicate are simply juxtaposed to the subject NP:

(20) TURKISH
Sevgi mühendis / ev-de
Sevgi engineer / house-LOC
np.h:s np:pred / np:l:pred
'Sevgi is an engineer / at home'

In such cases then annotators simply have to reach a decision on which element is the predicate, and what its form category is, then gloss accordingly. In examples such as the preceding one, this is quite straightforward. Problems may arise, however, when the predicate element carries some kind of verbal inflection (e.g. person agreement, or tense). To avoid undue complications at this point, we recommend that annotators ignore bound morphological expression of non-verbal predication, and gloss as in (20).

2.5.4 Existential and presentational predicates

Languages often have specialised clause constructions for the purpose of establishing a referent as a topic, or merely introduce it into the universe of discourse (cf.Lambrecht 1994:177f.). The English *there is* construction is an example:

(21) *There are* *cockroaches [...]*
other other:predex np:s

Lambrecht 1994:179

Despite the finite copula verb involved in this construction, we propose to not further analyse them into there constituent parts but gloss the combination of adverb and copula as a predicate of the category 'other'. This construction has special structural as well as functional properties: first it shows inverted word order, and also has a possible mismatch in number agreement. Also, as Lambrecht (1994:179)) points out, it has solely the function of stating the existence of the entity referred to by the subject NP. The gloss <other:predex> is thus used here for the entire expression *there was*, and the NP referring to the entity the existence of which is asserted receives S function. The same gloss is applied to existential particle construction, as found in Russian or Vera'a:

(22) a. RUSSIAN
bibliotek-a *est'*
library-NOM.SG.FEM exist
np:s other:predex
'There *is* a library.'

b. VERA'A
raes bēne?
rice exist
np:s other:predex
'Is there rice?' (i.e. 'Do you have rice (for sale)?')

Languages may in addition possess constructions that share some properties with a clearly specialised existential construction, but which nevertheless appear to be more similar to canonical finite clause constructions, for example in English:

(23) *Here comes the lads*
a. # other:predex ln np.h:s
b. # other v:pred ln np.h:s

This construction has apparently more a presentational rather than existential function, and it involves a canonical lexical verb rather than some kind of copula. The construction does involve subject inversion and there is a mismatch in number values between verb form and subject NP. It is thus both functionally and structurally different from a clause *The lads come here.* For such borderline cases, we offer the alternative annotations shown above, the one in (23a), reflecting more closely the presentational character of the construction, and (23b), the more canonical properties. As with certain types of copulas (cf. Section 2.5.2 above), annotators will have to decide whether a construction of this or similar type is best considered a 'normal' finite predicate construction, or a specialised existential/presentational construction that is best glossed as a whole, in analogy to the English examples presented here.

In yet other languages, existential and/or presentational constructions involve non-verbal NP predicates, the entity whose existence is asserted is the one referred to by this predicative NP. A typical example is the following from Vera'a, representing the canonical beginning of a customary story:

(24) *qōn̄* *ne* *vōwal,* *e* *ruwa* *mē* *=n* *gunu-ruō*
night NUM.ART one PERS.ART two.people DAT =ART spouse-3DL
np:other rn rn ln np.h:predex rn rn rn
'One day, (there were) two who were a couple.' ANV.001

Note that in this example there is no element in <s> function. When comparing Vera'a to other languages, then, it would be necessary to count instances of <np:predex> together with <np:s>.

2.5.5 Non-canonical predicates

Many languages predicates that are 'deverbalised' in various ways, yet may express propositions like regular verbal clause constructions. These constructions involve predicative forms that are not fully-fledged finite verb forms or finite verbal phrases, in that they often admit only a subset of the available TAM (and sometimes voice) distinctions in the language concerned. Typical examples of such constructions are converb or participial constructions, head-tail linkages, other infinitival constructions, and various types of nominalisation constructions (or 'Action Nominal Constructions' in terms of Koptjevskaja-Tamm 1993). The predicative form of these constructions is usually a deverbal form that basically 'inherits' the argument structure properties of the verb it is derived from, but, crucially for GRAID, often shows reduced possibilities for expressing verbal arguments, in particular S and A. In some languages, quite a large proportion of predicates in actual texts are carried by such forms, which raise certain problems for annotators.

It is simply not possible to cover all the attested types of non-finite, or less-finite clause constructions attested cross-linguistically in this manual. Our general solution is to use the gloss <vother> for those predicative elements which functionally fulfill a role similar to a canonical finite predicative form, but are deficient with regard to government of verbal arguments. The <vother>

gloss is combined with the <pred> gloss for function, yielding <vother:pred> when the form in question is considered to head a clause unit.

A more serious problem arises with the glossing of the unexpressed verbal arguments in such constructions. As mentioned above, we largely follow the maxime of Bickel (2003), that if a predicate systematically excludes the possibility of expressing S or A, then we do not include a <0>-gloss in the glossing. The <vother> gloss is basically a signal to be read as: "this is a predicate, but it does not necessarily require an S or A argument". In the quantitative analysis of the glossing, this can be crucial information, which needs to be considered when assessing, for example, the overall frequencies of arguments in a text. For some types of analysis, the analyst might in fact choose to ignore all clauses containing a <vother>. But we definitely recommend noting this information in the gloss, as it may be of considerable relevance in the overall profile of the language.

Note that some non-finite predicates which would be candidates for the <vother> gloss do in fact permit the expression of S or A arguments, but they are not marked in the normal way (they might be in a genitive case, for example). In such cases they may be glossed with the function-gloss <:ncs>. For other arguments, either <:obl>, or <:other> may be used. These points often apply to imperatives, which will often not allow any overt expression of S or A, and hence can be glossed <vother:pred>.

The <vother> gloss is also a useful option for verbal derivations that are used as complements to verbs such as 'stop', 'start', 'dislike' etc. For example, the sentence *Mary stopped / started / disliked drinking whiskey* could be glossed as follows: <np.h:a v:pred #cc:p vother np:p>.

2.6 Clause boundaries, embedded clauses, and clausal operators

Clause boundaries are marked by special symbols at the beginning of a clause, and in special instances of centre-embedded clauses with a symbol at the end of the clause. We discuss both cases in turn, and then provide a set of guidelines for annotators. Table 5 provides the symbols involved in the annotation of clause boundaries.

Table 5: Glosses for clause boundaries, embedded clauses, and clausal operators

##	boundary of independent clause, inserted at left edge
#	boundary of dependent clause, inserted at left edge, further specified
rc	relative clause
cc	complement clause
ac	adverbial clause
ds	direct speech
neg	negative polarity
%	end of a dependent clause (if not coinciding with the end if its main clause)

2.6.1 Boundaries of independent and dependent clauses

The main distinction drawn in GRAID is between main, i.e. syntactically independent, clauses, and those considered to be syntactically dependent on another

clause:

- The beginning of a main clause is indicated by a double hash: <##>
- The beginning of a dependent clause indicated by a single hash: <#>

A problem that arises with this basic distinction concerns various kinds of clause coordination ('and' etc.), which may just involve strings of asyndetic clauses with no overt coordinator. It has been pointed out many times in the literature that the two clauses linked by an apparently 'neutral' coordinator are often not syntactically equivalent, and that there is often a degree of dependence between them. It is therefore questionable whether both should be treated as main clauses, and our recommendation would be to insert the main clause gloss only at the beginnings of those clauses that appear genuinely independent, cf. the following paragraph for a possible solution.

Dependent clauses usually combine a single hash with a gloss for their function, e.g. <#rc> marks the beginning of a dependent clause functioning as a NP-level modifier, i.e. a relative clause. The symbol <#> can also be used on own in cases where the annotator wishes to leave the function unspecified. Single <#> can also be used for coordinated clauses which may show properties of main clauses but are nevertheless not fully independent, as outlined above. Table 5 provides glosses and related functions. The symbol <ds> is added to either <##> or <#> to signal that the clause constitutes direct speech. It has its own slot within the GRAID gloss word and can combinie with all other function symbols, cf. Table 6 below. The rationale behind this practice is the following: initial clauses of direct speech often look like some kind of complement clause constructions licensed by a *verbum dicendi* in the matrix clause, often being 'accomodated' by a particular marker, e.g. a quotative particle, that may be similar to a complementiser in the language concerned (cf. e.g. Vera'a, documentation). Subsequent clauses in direct speech, however, usually look basically like other main or embedded clauses in non-direct-speech discourse, hence analysing them as complement clauses appears to be inadequate. Nevertheless, direct speech can be expected to yield quite different properties in terms of types of referents and their formal expression involved, hence we recommend that they be marked off to allow for distinct anlaysis.

Among clausal operators, we consider only polarity. Both independent and dependent negated clauses are glossed with <neg>, yielding <##neg> for independent clauses. In dependent clauses, the negation symbol is combined with the relevant clause type symbol, yielding for instance <#rc.neg>, <#cc.neg>, etc. Affirmative clauses are not marked.

The syntactic function of dependent clauses can be glossed in the same way as that of NPs, for instance, <#cc:p>. Relative clauses that are attributes to nouns do not receive a function gloss. However, free or headless relatives that take on argument positions can be glossed for function, for instance, <#rc:s> in English *Who dares wins.*. It should be noted here that the more detailed glossing of dependent clauses is optional in GRAID. As the main focus is on NPs functioning as arguments, one may wish to consider only the arguments within embedded clauses and neglect the syntactic function of the entire clause within the matrix clause. This issue is taken up in Section 4.4 below.

2.6.2 Centre embeddings

One problem that may occur in different languages is the centre-embedding of different types of dependent clauses. For instance in head-final languages, complement clauses are necessarily centre-embedded, thus splitting, for instance, the subject of a clause from its predicate. Schematically this can be illustrated as follows, where square brackets enclose the center-embedded clause:

(25) # x:A [#cc:P ... v:pred] v:pred # ...

For such cases, neglecting the hierarchical structure and simply glossing along linear order of NPs and predicates would create the wrong impression that there is a clause unit containing only an <A> argument, and it would also leave two instances of <pred> in the second chunk.

Our solution for GRAID annotations is as follows: if a dependent clause is centre-embedded in the main clause, or the main clause starts with an embedded clause, i.e. the main clause continues after the end of the dependent clause, then the end of the dependent clause is marked with <%>. If an embedded clause is followed immediately by another embedded clause, the end of the first one is not marked by <%>. Thus:

- The symbol % is used as a right-boundary symbol of an embedded clause if it does not coincide with
  a. the end of the clause in which it is embedded or
  b. the beginning of another embedded clause.

2.6.3 Overview of clause-boundary glosses

In addition to the above, various kinds of information can be included with the clause-boundary gloss. The structure of a clause boundary gloss is summed up in the following schema, though not all possible combinations will be actually attested:

Table 6: Possible clause boundary gloss structures

#	ds	_	(default: empty)	.	neg	:	p
##			cc				pred
			ac				other
			rc				obl

2.6.4 Instructions for glossing

The most basic level of glossing would distinguish main from dependent clauses, and would also note direct speech and negation, but make no further distinctions. A higher level of detail would distinguish between different kinds of dependent clause (rc, cc, ac etc.), and would also note syntactic function where necessary. In the case of headless relatives, it can be argued that the clause is a referential item, and therefore could be given a modifier such as <h> 'human referent'.

##	leftward-boundary of main clause
#	leftward-boundary of a dependent clause, type and function not specified
#ds	dependent clause, direct speech, not negated,
##ds	independent clause, direct speech, not negated function not specified
#ds_rc	relative clause rendering direct speech
#ac.neg	adverbial clause, negated
#ds_cc.neg:p	complement clause of a transitive verb, negated, rendering direct speech

The handling of direct speech in GRAID needs some comments. Firstly, direct speech as occurring in narratives complements verbs of saying, and is thus part of complex sentences. As such, it may be in paradigmatic relation to content-denoting NPs, e.g. in English *He often says stupid things that upset her.* and *He often says: 'I hate our neighbours!'*. On the other hand, direct speech may occur without any ntroduction by a verb of saying, thus not being embedded in a complex sentence structure. This is essentially also true in cases of longer stretches of direct speech where probably only its first clause can be considered a genuine complement in the matrix clause. Therefore, direct speech is not considered here as paradigmatically related to complemented and other embedded clauses, the <ds> symbol thus being compatible with the main clause boundary symbol <##>, yielding <##ds>. Moreover, direct speech may of course show complex sentences that contain dependent clauses, as exemplified in the table above. In instances of inner-direct-speech complement or relative clauses, the single <#> is used. Obviously then, complementation by direct speech differs considerably from complementation by a NP with P function, which raises issues for the treatment of argument functions in these constructions, to be discussed in 4.4 below.

Example sentences from ENGLISH:

(26) *the guy we talked about is coming later.*
ln np.h:s #rc pro.1:s v:pred rv % aux v:pred other

(27) *I'm not very happy about it.*
##neg pro.1:s=cop lv lv other:pred adp pro:obl

(28) *That he came to the party surprised me.*
#cc:a other pro.h:s v:pred adp ln np:g % v:pred pro.1:p

Examples from BEAVER:

(29) *eskee ʔíge kwehnushi? kudyi*
np.h:a rn #ds$_c$c:p pro.1:s-v:pred % pro:p-0.h:a-v:pred
'A young man wants to get married.' anecdote marriage 001:1

(30) *dáá wǫhch'e ǫ ghajii*
#ds vother nc % pl_pro.h:s-v:pred
"What happened?' they asked.' Snare Hill001:72

(31) *ts'ído as?idǫ keenasdyíhó ...*
#ac np.h:pred pro.1:s-cop % pro:p-pro.1:a-v:pred
'When I was a child, I remember...' hunt ducklings:1

(32) *kéénasdyehe* *tye'a* *gǫ* *kataa'ayaa* *e?a*
pro:p-pro.1:a-v:pred #cc:p np.h:s other pro:obl-pro.h:s-v:pred other
'I remember that my dad used to go along with them' hunt ducklings:9

(33) *dyuhdyee, dyuhdyee, kudzi?, kudzi?* *dáághadyishu*
#ds other other other other % pro.h.pl:s-v:pred
"Way over here, over here, over there, over there' they used to say'
hunt ducklings:18

2.7 'Non-classifiable', and 'other'

Annotators will usually come across two types of element that can not be handled with the glossing aparatus outlined thus far. For one thing, it is not always possible to reach a principled decision on how to gloss a given object language element. Other elements do not pose any analytical problmes, but are simply not relevant here because they do not impact on the issues GRAID is designed to tackle. In such cases, GRAID offers two respective options, shown in Table 7.

Table 7: Glosses for irrelevant and non-classifiable elements

other	forms / words / elements which are not relevant for the analysis
nc	'not considered' / 'non-classifiable'

The <other> gloss is primarily used for elements which are outside the purview of grammatical relations in the narrow sense, for example various types of adverbs, interjections, interrogative particles, or discourse particles. This gloss can also be used for elements that appear to fulfill an argument function, e.g. locatives, but cannot be unambiguously assigned a form category such as pronoun or NP. This is the case with certain types of local adverbs, e.g. 'inside', or for the object of a verb of speech, as in *he said "hey!"*. The word *hey!* can be considered an object to *said*, but it would be difficult to classify it as a NP, or even a complement clause. Thus we would recommend here <other:p>.

The gloss <other> can also be used in both the function slot, i.e. <np:other>, for example with NPs which express circumstantials rather than discourse participants, or with the non-verbal complement of a copular verb, cf. (19) above). The <other> gloss can also be used for forms, for example interrogative pronouns (*who, what* etc.), which have a discourse and reference function very different from other pronoun types, e.g. they are not obviously anaphoric. Annotators may therefore decide to gloss them as <other>, rather than make a choice between <np> and <pro>, while still assigning them an unambiguous function gloss (e.g. <other:P> for *who* in *who did you see?*). Where both the form and the function of an expression are classifed as 'other', we suggest simoply annotating the expression as <other> rather than <other:other>.

The gloss <nc> is intended to be used where annotators are not sure how to analyze a particular expression or construction due to the following kinds of difficulties:

- The analysis of a given construction remains unclear at a given stage of investigation.

- The construction is incomplete, thus not a valid construction in the given context (false starts, interruptions, or parts are inaudible).
- The words or construction under consideration constitute a formulaic expression displaying a highly idiosyncratic syntax, or one that is not amenable to conventional analysis in terms of predicate/argument structure.
- Phrasal interjections, hedges, rhetorical devices (cf. English *you know, I mean, right?* etc.) which may be prosodically independent utterances, but lack obvious argument-predicate relations.
- The annotator may choose to systematically ignore a particular construction type that is, for example, rare in the corpus and would otherwise pose considerable difficulties in glossing.
- Recorded stretches of discourse that are not transcribed due to various reasons (not intelligible, not audible) and hence given as 'non-audible' in the transcription tier will also receive an <nc> gloss in the GRAID annotation.

As in all contentious issues in glossing, the choice of solutions will depend to a large extent on the **relative frequency of the problem cases in texts**. For example, if some non-finite verb form only occurs perhaps once in 200 clause units, it is simply not necessary to spend time working out a specific glossing solution; the clause concerned can simply be glossed <nc>. If, on the other hand, such forms are quite frequent, say 10-20 examples in 200 clause units, annotators need to decide on a consistent treatment, note it in the documentation, and adhere to it consistently. Our initial experience with Gorani, Vera'a and various other languages is that roughly 10% of the clause units in a given text are <nc>. This appears to be a tolerable level; should the number of <nc>-units rise significantly above this, the annotator may need to reconsider some of the glossing solutions.

2.8 Annotation of phrase-internal constituents

As repeatedly stated above, GRAID glosses primarily target clause-level constituents. These are NPs, VPs (in the narrow sense, excluding NPs, PPs, etc.), adpositional and other phrases. In some languages, a phrase often coincides with a single word form to which the respective GRAID gloss can be readily applied. In other languages, however, expressions for arguments and predicates are multi-word phrases by default. In the following, we provide a set of rules for the glossing of multi-word phrases.

Obviously, as GRAID glossing takes note of the form and function of clause-level constituents, the question arises as to which word of, e.g. a complex NP, should be aligned with the gloss. In general, we take the lexical head as the locus for the main GRAID gloss. Glossing only the lexical heads of clause-level constituents is the most basic level of annotation detail. This can then be refined in a step-wise model of annotation detail. Consider the following:

A: Basic annotation

	those	*crazy*	*linguists*	*are working hard*
#			np.h:s	aux v:pred

In this example, the head of the NP and the head of the VP take the GRAID gloss. Other constituents are simply left unglossed. This is the most basic level of detail for GRAID annotations, and it allows for quantitative analysis of how many NPs (as opposed to pronouns, zeroes, etc.) are in which functions, with what type of referent; but it does not provide information about degrees of complexity (or 'weight') of different NPs.

A further layer of annotation detail would take note of subconstituents of a NP or VP:

B: Subconstituent annotation

	those	*crazy*	*linguists*	*are*	*working*	*hard*
#	ln	ln	np.h:s	aux	v:pred	rv

Here, the glosses <ln> stand for 'constituent of a NP, appearing to the left of its lexical head', and—for <rv>—'constituent of a VP, appearing to the right of its head'. Note that we do not suggest a neat rendering of hierarchical structures, but consider only the linear order of elements relative to the phrasal head on the topmost layer of phrase-internal structure. The 'Subconstituent annotation' of this type enables phrase boundaries to be identified, and provides an indication of the relative complexity / weight of different phrases and phrase types. It could then be used to investigate, for instance, the functional distribution of NPs of different weight. We do not distinguish the class membership of different elements on this level of annotation detail, hence that the first element of the NP is a demonstrative and the second an adjective, is neglected.

In many isolating languages, e.g. Oceanic languages, expression of arguments and predicates by multiple-word phrases will be the rule rather than a special case. Hence, the example (3) from Vera'a mentioned above can alternatively be glossed as follows, considering all the subconstituents:

(34)

	[...]	***ne***	***kal***	***'ō'***	***ba'a***	***kēl***	***sar***	*ēn*	*'añsara*	*ē*
		TAM2:3SG	enter	carry	inside	back	in	ART	person	DEM3
#	0.h:a	lv	v:pred	rv	rv	rv	rv	ln	np.h:p	rn
lē	*=n*	*mē'ērsa*								
LOC	=ART	harbour								
adp	ln	np:g								

'... and then took that man back ashore at a harbour.' ISAM.065

This second level of annotation is exemplified in examples in the preceding sections.

A third layer of detail could potentially take note of category membership, as shown in the following example:

C: Lexical category annotation

	those	*crazy*	*linguists*	*are*	*working*	*hard*
#	ln_dem	ln_adj	np.h:s	aux	v:pred	rv_adj

Where lexical category labels are employed, these would simply be added to the left/right-ordering symbols used at Level B, essentially in the way all additional information is added to GRAID annotation, as outlined in 2.9 below. The categories identified must of course be defined in the documentation, together with the labels employed. This type of very fine-grained annotation will obviously

only be employed by researchers who work on very specific research questions requiring this level of detail. Researchers may of course decide to add only those specific category labels that are relevan for the particular research they are engaged in, e.g. taking note only of the presence of demonstratives:

C': Partial lexical category annotation

	those	*crazy*	*linguists*	*are*	*working*	*hard*
#	ln_dem	ln	np.h:s	aux	v:pred	rv

This latter type of annotation detail could for instance be used to engage in research about the functional and discourse distribution of NPs containing demonstratives. Which level of granularity is adopted depends not only on the additional research questions of individual annotators, but also on practical considerations such as the intended speed of annotation, available resources for undertaking these annotations, etc. It is fairly obvious that annotations on Level A can be undertaken much quicker than those on Level B or Level C. But note that in the quantitative analysis, the finer-grained analysis should be in principle comparable with the coarse-grained one, at least to the level of granularity achieved by the coarser-grained analysis. In search queries, then, only the basic GRAID symbols might be considered, and the others would simply be ignored. It is a very important principle of GRAID that a finer-grained analysis, if intended, still be compatible with the basic level analysis.

2.9 Adding further detail to GRAID annotations

GRAID has been designed for very specific research issues, and hence many grammatical and semantic distinctions encoded in natural languages have been left unconsidered for our purposes. Moreover, we have tried to keep the inventory of glosses as small as possible to foster practicality of glossing. After all, annotations of the kind proposed here are potentially quite resource intensive, hence there is naturally a strong motivation to keep things to a minimum.

A number of grammatical categories and semantic properties of referential expressions may nevertheless be worth noting, at least in some languages. For these purposes, annotators may wish to introduce language-specific tags over and above the core inventory outlined in this manual. An example is number distinctions in referential expressions which may indeed prove to be quite relevant for the formal expression of arguments. If we consider for now only the distinction between singular and plural, this can be noted as follows:

(35) a.

	those	*crazy*	*linguists*	*are*	*working*	*hard*
#	ln	ln	pl_np.h:s	aux	v:pred	rv

b.

	that	*crazy*	*linguist*	*is*	*working*	*hard*
#	ln	ln	np.h:s	aux	v:pred	rv

If additional tags are used, we urge that the following rules be adhered to: annotators should not change or impair the original GRAD gloss words. Thus, additional glosses should be added at the margins of a GRAID gloss word, separated by an underscore, so that the original gloss can still be easily searched for and analysed. We therefore suggest that additions to the form gloss be attached to the left, while additions to the function gloss be attached to the right. Also, glossing should be kept economic, so that one value of a

particular category—preferably the unmarked and most frequent one—be kept unglossed, as is done with the singular in example (35). Finally, annotators need to document which categories they have added and how they are glossesd, including what value is indicated by 'zero-gloss'.

By the same token, annotators may also wish to note further subclasses of, for instance, the referential form classes proposed here. An example in place are demonstrative pronouns which could be glossed as follows if annotators wish to preserve this information:

(36) *Those* *are expensive.*
dem_pro.h:s cop other:pred

Again, such practices will need to be documented by annotators for the respective annotated corpus.

3 Argument indexing and agreement

One of the central research issues that can be addressed using GRAID concerns the way participants in events receive linguistic expression in actual discourse. Reference is typically effected through different types of referential expressions, like lexical NPs, personal pronouns, or zero anaphora. Non-lexical forms of expression come in various forms, which Siewierska (2004) collectively labels 'person markers'. These include free pronouns, clitic or prosodically defective pronouns, or affixes. In the literature, a distinction is traditionally drawn between "agreement" on the one hand, and "pronominal", or "anaphoric" uses of person markers on the other (the terminology is notoriously inconsistent in this area; our use of "agreement" here is the more traditional one, not compatible with Siewierska's extended use of the term). The basic insight behind this distinction is that "agreement" is considered to be a more or less mechanical replication of certain features (for example, person and number), triggered by the presence of the actual argument of the verb, while pronominal person markers are considered to represent the arguments themselves. However, there is no consensus on the criteria by which the distinction is drawn, and controversy regarding the correct analysis of even well-researched languages such as Spanish continues to rage. We share Haspelmath's (2013) view that the anaphora/agreement distinction is almost entirely motivated by theory-internal considerations (''functional uniqueness" of *Lexical Functional Grammar* etc., cf. relevant passages in Bresnan and McHombo (1987) and Bresnan (2001)), and there is in fact no necessity to apply the dichotomy when investigating person markers in discourse. We discuss Haspelmath's (2013) and Corbett's (2003) approaches in the next section below.

The standpoint adopted here is that reference in discourse is an empirical issue, and the formatives involved should—initially at least—be taken at face value, with a minimum of theoretical pre-judgement. Whether the dichotomy is reflected in a significant manner in discourse is a question that can only be meaningfully addressed after a significant amount of actual data has been analysed.

In what follows we will first outline some practical considerations for GRAID annotators, and then sum up the theoretical discussion of argument indexing

and agreement, and how GRAID-annotated corpora can in fact contribute to this line of research.

3.1 Practical outline for annotators

The overarching principle behind annotating person markers in GRAID is that we annotate those positions which **permit variation**, that is, where the presence or absence of a person marker is not fully predictable, but is (co-)determined by non-syntactic factors (stylistic factors, or considerations of information management, for example). Before turning to the ramifications of this principle, we will introduce the different possibilities for annotating the **form** of a person marker. GRAID recognizes three possibilities:

1. Free pronoun. A person marker that is capable of bearing independent stress and has (some degree of) syntactic mobility. For example, the pronoun *I* in English is considered a free pronoun. Although its degree of syntactic mobility is very restricted (it can only occur before a finite verb), it can still be separated from the finite verb by certain adverbs, as in (*I really don't like turnips.* A free pronoun such as *I* in English is glossed <pro.1>[7].

2. Clitic pronoun. A person marker that lacks independent stress, and is thus prosodically dependent on another word as host, but which is not strictly subcategorized for the category of its host. Clitic pronouns are glossed <=pro> (enclitic), or <pro=> (proclitic).

3. Affix. A person marker which is (i) prosodically dependent (bound), (ii) strictly subcategorized for a particular host (for instance verbs), and (iii) exhibits the formal properties of other inflectional affixes in the language concerned (with respect to for example vowel harmony, or morphophonological processes etc.) can be considered an affix. Affixal person markers are annotated with <-pro> (suffix) or <pro-> (prefix).

Decisions on what is to be considered a clitic and what an affix, can only be reached after consideration of the language-specific morpho-syntax, and need to be briefly justified in the accompanying documentation.

As a general rule, free pronouns are **always** glossed in GRAID annotations, while clitics or affixes may not be, depending on the degree of obligatoriness governing their realization. This is the issue we now turn to.

Before beginning annotation, investigators need to consider whether a particular person marker occurs obligatorily and is thus present in all instances of a particular argument function due to a categorical rule (cf. Haspelmath (2013) 'gram-index' and many 'cross-index' systems discussed in Section 3.2 below). As a rule, these person forms do not need to be glossed at all in GRAID for the simple reason that their occurrence is categorical, hence no variation can be expected to be found.

A straightforward example of such categorically obligatory person markers is affixal subject agreement in many languages, for example German. Every finite

[7] Optionally, one might wish to distinguish between *me* and *I*, because the former has greater syntactic freedom. This could be achieved with the <w> 'weak' tag, as in <wpro> vs. <pro>.

verb in German carries a marker indexing the person of the subject (S or A). The presence of these markers is not subject to pragmatically-driven variation, and they are therefore quite predictably present. Our recommendation is therefore not to annotate these person markers, as their presence can be inferred via a general rule, and can simply be factored into a quantitative analysis by counting the numbers of finite verbs.

A more complex example of obligatory bound indexing comes from Sakapultek Maya (data from Du Bois 1987). Sakapultek has ergative alignment and two sets of bound person indexes (i.e. affixes) occurring in different morphological slots of the verb. Indexes for S, A and P arguments of the 1st and 2nd person are obligatory:

(37) SAKAPULTEK

a. *š-**at-qa**-kuna-:x*
TAM-2SG.ABS-1PL.ERG-cure-TR
'**We** cured **you (sg)**.' Du Bois 1987:809

b. *š-**ax-a:**-kuna-:x*
TAM-1PL.ABS-2SG.ERG-cure-TR
'**You (sg)** cured **us**.' Du Bois 1987:809

c. *š-**ax**-war-ek*
TAM-1pl.abs -sleep-ITR
'**We** slept.' Du Bois 1987:810

d. *e: ra ax k-**ax**-war-ek*
FOC the 1PL TAM-1PL.ABS-sleep-ITR
'**We** slept.' (or: 'It was US who slept.') Du Bois 1987:810

The last example shows that 1st and 2nd person indexes occur regardless of whether a free expression of the same referent is co-present on clause level. Du Bois explains that in such cases, they are additionally accompanied by a focus marker and a determiner, hence *e: ra ax* 'we'. As Du Bois (1987:810) states, the occurrence of such free pronominal expressions is extremely rare and pragmatically restricted to contrastive contexts (cf. Yup'ik discussed below). The situation with 3rd person arguments is similar:

(38) SAKAPULTEK

a. *k-0-a:-kuna-:x*
TAM-3.ABS-2SG.ERG-cure-TR
'You (sg) cure him.'

b. *k-0-war-ek*
TAM-3.ABS-sleep-ITR
'He sleeps.'

c. *k-0-war l ačen*
TAM-3.ABS-sleep the man
'The man sleeps.'

In Sakapultek, then, we have invariable occurrence of a bound (affixal) person index in one position for each argument function (i.e. on the verb). This contrasts with the situation for free expressions, where we find variation between

NP, free pronominal expression (though this is rare, it is still apparently possible) and zero. Given that here we find variation, it is necessary to annotate accordingly, so the GRAID-annotation of 38a through 38c would be as follows:

1. <0.2:a 0.h:p v:pred>
2. <0.h:s v:pred>
3. <v:pred np.h:s>

Now given the paucity of free pronouns in Sakapultek, we can expect to find a high number of zeros in the GRAID glossing (at least significantly higher than in English), and it is precisely this result which gives us a quantifiable measure of the actual difference between the two languages. This is quite a different approach to one in terms of pre-fabricated categories, such as "head-marking" vs. "dependent-marking", or "pro-drop" versus "non-pro-drop" etc. Although a GRAID annotation of Sakapultek along the lines just suggested would yield a very high number of zeroes, it would not be 100% of S and A's, simply because we still find significant numbers of full NPs for S and A. Just how few overt arguments are expressed in Sakapultek discourse, and how it compares to other languages, is an open question that needs to be tackled in a GRAID annotation, rather than merely assumed a priori. With regard to the presence of the affixal person markers on the verb, this is captured in the general documentation accompaning the annotation, and can be factored into the investigation when considering the (possible) impact of these bound forms ("mentions" in DuBois' terminology) on the density of overt forms in discourse. GRAID annotations thus do not prejudge what position, verb index or free form, should be considered the 'real' argument expression, but are instead intended to feed into investigations of this sort.

Similarly unproblematic for GRAID annotations are those systems where a bound person index is clearly in complementary distribution with a free exponent of an argumemt role, so that the two types of expression are mutually exclusive. The following examples illustrate such a indexing sysem in Central Kurdish (Indo-European, Iranian; North Iraq). In Central Kurdish, all finite verbs in the present tenses carry canonical affixal agreement with S/A, which will not be glossed. But in addition, there is also a set of person clitics that may be used for various syntactic functions, for example direct object, or prepositional complement. When expressing a direct object with present-tense verbs, the clitic is in complementary distribution with a full pronoun or NP: if the latter is overt, then there is no clitic pronoun (39a). If the latter is not present, the corresponding clitic pronoun attaches to the left-most constituent of the VP, in (39b) the negation prefix:

(39) Central Kurdish

a. *Min to na-bîn-im*
   1S 2S NEG-see:PRES-1S
   'I don't see you'

b. *Min na=t=bîn-im*
   1S NEG=2S=see:PRES-1S
   'I don't see you'

But it is ungrammatical to have both the full pronoun and the clitic in the clause, as shown in (40):

(40) **Min to na=t=bîn-im*
1S 2S NEG=2S=see:PRES-1S

In Haspelmath's (2013) terminology, the clitic object pronoun is a pro-index (cf. Section 3.2 below): they are not obligatory (not required in all clauses of this type), but only when an overt NP or free pronoun object is not present. For the GRAID annotation, it follows that in Sakapultek an object index on the verb would not be glossed, while an object index in Central Kurdish needs to be glossed, with an equal-sign boundary symbol indicating the fact that the form is a clitic: <=pro>.

In systems of so-called 'pro-indexing' (cf. Section 3.2 below), where NP and bound person index cannot co-occur, we recommend glossing them as alternate forms. Thus if the bound pronoun occurs, we gloss it, but do not gloss a zero in the clause. A hyphen or equal sign respectively attached to <pro> will preserve the information that a "pronominal" form is, in terms of its realization properties, an affix or a clitic.

We should point out that the facts from Central Kurdish just discussed hold only for transitive verbs in the present tense. In past tenses, clitic deployment is subject to quite different rules, which we will not discuss here (cf. Haig (2008:Ch. 6) details). But the point is that annotators should be aware of the fact that argument indexing systems are often construction-specific, rather than language-specific, and annotators may well need to define their annotation processes for distinct constructions, and if necessary add language-specific tags to enable distinct subsystems to be identified.

Potentially problematic for GRAID are those cases of indexing systems that have both free and bound person markers, and allow both to co-occur within the same clause, but do not require them to do so. In other words, both sets are subject to variation. Let us consider the example of pronominalization of RECIPIENTs in Spanish. Spanish shows what has traditionally been called 'clitic pronouns' (data from Pineda and Meza, Undated), illustrated in the following examples:

(41) SPANISH
Juan muestra el catálogo a Maria
J. show:PST:3S the catalogue to Maria
'Juan showed the catalogue to Maria'

Depending on the larger discourse context and the communicative intentions of the speaker, both the direct as well as the indirect objects of this sentence can be pronominalized, using pre- or postclitics. The following constellations of NPs and pronouns (among others) are possible:

(42) SPANISH
a. *Muestra=lo a Maria*
b. *Lo=muestra a Maria*
c. *Muestra=le el catálogo*
d. *se=lo=muestra*

e. *se=lo=muestra a Maria*

In the examples (42a) and (42b), only the THEME is pronominalized (the clitic *lo*) while the RECIPIENT remains as a full PP (*a Maria*). In (42c), on the other hand, the RECIPIENT is pronominalized (*le*), while the THEME is a full NP. In (42d), both the RECIPIENT (*se*) and the THEME (*lo*) have been pronominalized. The same is true of (42e), but here the RECIPIENT also occurs, seemingly redundantly, as a full PP in the clause (*a Maria*).

For Spanish, annotation of subject and direct objects in GRAID would be straightforward: for subjects, the verb shows regular person inflection, and thus only the free form of representation needs to be registered. For objects (at least inanimate ones), the clitic pronoun and NP are mutually exclusive, so only one needs to be glossed. Problematic is the representation of the RECIPIENT argument, as it shows variation in both free and bound pronouns: it may be represented by a full NP within the clause, or left unexpressed, and it may or may not have a clitic pronoun on the verb. Obviously, both sets of representation need to be registered in GRAID. Our recommendation in such cases is therefore to gloss both the clitic pronoun and the NP-argument, yielding for example the following:

(43) SPANISH
se=lo=muestra *a Maria*
3S:DAT=3S:ACC=show:PST:3S to Maria
pro.h:g=pro:p=v:pred adp np.h:g
'He showed it to Maria'

The same will apply to the postclitic in (42c). Where no bound index occurs, no glossing is applied, and the relative frequency of presence versus absence of of bound indexes for a particular argument role can be recovered later via a filtered search. One will still have to gloss zero forms for the RECIPIENT at clause level in examples (42c) and (42d).

While the glossing of such cases of multirepresentation may not be too problematic as such, the analysis of such multiple sets of GRAID glosses for a single argument role is challenging. The problem is that a global search for free and bound person markers, when run over an entire text, will simply yield raw figures for bound and free argument expressions, but will not tell us about where the two co-occur in the same clause. However, this can be achieved through more complex filtered searches and regular expressions, so that quantitative measures of co-occurrence of bound and free forms can be extracted (cf. Schnell and Haig (In print) for an example).

While the issue of multirepresentation and co-variance may be fairly marginal in Spanish, as it concerns only the fairly rare instances of RECIPIENT arguments, this may be quite central in other languages where both bound person indexes and free expressions for core argument roles S, A or P may be possible but not obligatory, hence leading to a massive degree of co-variance (e.g. in S and A arguments in Ingush; P arguments in Teop; etc.).The issue of co-variance may in fact also arise in languages where different verbal stems behave differently in terms of person marking. Well-known example are Semitic languages, like Modern Israeli (Ivrit), where present tense stems do not take person markers, but stems of other tenses do, e.g. past tense or future ones (cf. Ariel 2000).

Similar issues arise in Russian, where past tense verb forms are not inflected for person, but only for gender and number (with gender being neutralised in plural forms), or in fact some Kurdish languages, as indicated above. In a language like Modern Israeli then, we find obligatory person indexes for S and A arguments in past tense, and in principle these would not have to be glossed according to the rules outlined above. However, person-marked forms are likely to alternate with unmarked present tense forms within the same text, and not annotating verbal person marking would thus yield the wrong impression (cf. Ariel 2000 for discussion of 1st and 2nd person 'zero' exponence). Our recommendation in cases like Modern Israeli is therefore to gloss all cases of bound person markers and 'bound zeroes', again using hyphen and equal sign to indicate the boundedness of the form, essentially like the RECIPIENT argument in Spanish or P arguments in Teop. In such cases, annotators may wish to add an optional language-specific tag to the predicate glosses indicating present vs. past tenses.

The following is a list of recommendations which we will further comment on below:

- Clearly and consistently obligatory argument indexing need not be glossed.
- Where indexing is optional or conditioned (e.g. by animacy or definiteness), argument indexes should be glossed and absence noted via a zero.
- Where different constructions in the same language have different argument indexing properties (e.g. past versus present in Ivrit), annotators may wish to add a language-specific tag to the predicate gloss to identify verb forms from the respective paradigms (see below).

In some languages with basically obligatory argument indexing for particular argument roles, an index may be suspended under very specific pragmatic conditions. Thus, for instance in Yimas, indexing of S and A arguments is entirely grammaticalised, and yet no index appears in cases where the respective referent is new (cf. Foley 1991:232ff.). Similarly in Makassarese, S and A arguments are not indexed if indefinite (Jukes 2005:662). In cases where an index is absent in very particular conditions and obviously very rarely, annotators may choose to follow the first rule, treating indexing basically as consistently present, and noting the few cases where it is actually absent. Whether the absence of indexing is restricted to rare cases of suspension, or indexing is in fact optional or conditioned, is a question an annotator will need to decide for a given language, and the decision can often only be taken after pilot annotation of around 100 clauses has been conducted; this provides the annotator with a rough benchmark of how frequent the "odd" cases are. Only where these cases occur sufficiently frequently will it be necessary to adapt the annotation. In cases of doubt, we recommend following the second rule, and note the absence versus presence of indexing by glossing <pro> vs. <0> (zero).

With regards to the third point, it may be an interesting research question to compare rates of overt versus non-overt arguments in clauses with different types of verb exhibiting different types of argument indexing (an investigation of this type for Russian present tense clauses versus past tense clauses is discussed in Kibrik 2011). If that is a focus of interest, annotators will need to identify the different verb forms; this can be done through an additional language-specific tag

added to the left of a predicate gloss, e.g. <pst_v:pred> for past tense verb forms (but note that this information may already be contained in the morphological glossing (if present) and could thus be recovered semi-automatically from that tier, thus avoiding the necessity for additional tags in the GRAID-tier). The alternative approach is to consider the different constructions as part of the overall variation in argument indexing, and gloss it according to the second rule above, regarding the entire verbal system as showing variable indexing.

Annotators should also be aware that there are likely to be differences in the way the language treats argument indexes for the first and second person markers, and how it treats third person forms.

Two further issues concern sets of person markers that fuse person/number indexing for more than one role, and constructional variation conditioned by definiteness of an argument. An example is the polysynthetic language Yup'ik, discussed in Mithun (2003). Note that in reality the Yup'ik operson markers would not need to be glossed at all due to their categorical occurrence. However, it is a good example for illustrating the treatment of fused person/number indexes. In Yup'ik, a single index set is used for person and number values of two argument roles at the same time:

(44) a. *Kassuutellrua.*
kassuute-llru-a-a
marry-PAST-TRANSITIVE.INDICATIVE-3SG/3SG
'He married her.' or 'She married him.'
the reading 'He married someone' is not available.

b. *Kassuutellruuq.*
kassuute-llru-u-q
marry-PAST-TRANSITIVE.INDICATIVE-3SG
'He got married.'

In this pair of examples, the word-final suffixes *-a* and *-q* represent participant roles. However, the *-a* suffix does not index a distinct referent's roles, but instead seems to simply indicate that two distinct specific referents are involved in a reciprocal state-of-affairs. This situation can be treated in GRAID as if the sufix fuses indexes for two argument roles with two participant referents, as is basically also done in the morpheme glossing provided by Mithun (2003), thus yielding a GRAID gloss like: <-pro.h:a/p>. While in this example, the two participants are equal in terms of person/number and animacy, the suffix *-put* in the following example represents two participant roles with different person values. Here, one would need to take note of this in the gloss: <-pro.1:a/h:p>:

(45) YUP'IK (Mithun 2003:243)

a. *arulaiqarlu**ta***
arula-ir-qar-lu-**ta**
be.in.motion-NEG-briefly-SUBORD-1PL
'... **we** stop briefly'

b. *nayuqaqurla**put***
nayur-qaqur-la-**put**
observe-intermittently-OPTATIVE-1PL/3PL
'and watch **them** for a while?'

Thus, the backslash signals that a single formative involves reference to two distinct participants, and it is applied at that point in the gloss at which the two deviate in terms of their role and reference properties, thus only between roles in the case of the suffix *-a*, and between person and roles in the case of *-put*. The suffixes *-q* and *-ta* index only one participant and can be treated like any other person index. Note, however, that under the assumption that bound argument indexing in Yup'ik, as in Sakapultek, is categorical and thus predictable, then the Yup'ik person markers would not need to be glossed at all. The Yup'ik example nevertheless illustrates how languages with fused indexes can be handled in GRAID. As for the clause-level position, we recommend glossing for NP, pro, or zero, despite the reasonable assumption that free pronouns in Yup'ik are very rare and pragmatically marked, for the reasons already outlined in connection with Sakapultek above.

The Yup'ik examples also show that bound person forms in Yup'ik do not allow for a non-specific reading of the P argument; where no specific P argument is involved, an alternative (intransitive) construction is employed, involving the single-role suffix (Yup'ik, Mithun 2003:251–252). In general according to Mithun (2003), Yup'ik person forms behave very much like English free object pronouns in that they occur only in contexts where their referent is identifiable. What is relevant for GRAID annotations is that the non-specific participant does not receive zero glossing, according to the rules outlined above. This would be so for the Yup'ik example in 45 as well as for the English equivalent.

The Yup'ik case also raises the issue of how to deal with systems which are sensitive to certain constellations of participants with regards to their relative ranking in terms of person and/or animacy, e.g. inverse systems in Algonquian and Na-Dene languages. We only give three general rules:

1. Clause-level expressions are glossed as NP / pro / zero.
2. Where the morphological marker is obligatory, it does not need to be glossed, and this is stated in the documentation for the language. It will be understood that every argument glossed on clause level will show additional representation in some form on the verb.
3. Where the marker is conditioned or otherwise readily omitted in various contexts, it will be glossed as shown here for Yup'ik.

Generally, we target person-based indexing as opposed to indexing based on other categories like gender, shape-classification, etc. We provide nothing in the core inventory to annotate such categories; annotators should consider individual solutions. Under this view, a gender-based indexing system as found in past tense forms of Russian verbs would be either not glossed, or annotators will need to introduce a language-specific tag for gender/number.

3.2 Theoretical approaches to argument indexing in discourse

In the preceding sections we considered some practical issues of glossing person forms. In this section we take up some of the theoretical discussion that has informed our annotation practice. As mentioned, there is an ongoing controversy in the literature with regard to the (non-)referential status of bound person

markers in some languages. However, reducing the issue to whether a particular person marker "is" or "is not" referential appears to be based on a very naïve conception of how reference is achieved in natural language: it is surely not the case that reference to participants in events is effected by just one particular marker, any more than reference to, for example, temporal setting, is carried by only one element in a clause (it might be carried by an adverb and a tense affix, for example). Rather, we assume that reference to a particular participant may be effected by several elements in the clause, that is, we explicitly endorse the notion of distributed information in the realm of reference. On this view, there is nothing intrinsically "wrong" with the idea that both a free pronoun and a same-clause, bound, co-referential person marker, contribute jointly to establishing reference. There is no necessity to consider one as the "real" argument, while the other is then by fiat necessarily something else (for example, just agreement). The question of how often such multiple representations occur, in which languages, and under which conditions, then emerges as an interesting research agenda, rather than something that has to be excluded from the outset by some notational convention. Recently, theoretical proposals along these lines have been formulated by Haspelmath (2013), which we take up here. Haspelmath introduces the term "index" as a cover term for different types of person forms that have traditionally been called 'cross-referencing', 'pronominal affix', or 'agreement'. Though we basically follow this terminology, we will also continue to refer to "agreement" in the sense of "canonical agreement", a notion developed in Corbett (2003, 2006), which we also discuss in this section[8].

Haspelmath (2013) distinguishes three types of systems involving the following types of argument indexes: 'gram-indexes', 'cross-indexes' and 'pro-indexes'. These types of indexing systems are defined exclusively on the criterion of 'conomination', i.e. whether a further (often, but not always, more informative) expression may co-occur with the argument index in the same 'narrow clause', thus excluding dislocated expressions[9]. According to this criterion then, 'gram-indexes' are defined as obligatory having conominals, 'cross-indexes' as being capable of having conominals, and 'pro-indexes' as not allowing for conominals. Gram-indexing is what in fact resembles 'agreement' in the more traditional sense most closely, as here the argument index obligatorily requires an overt antecedent in the same clause which can be analysed as the controller of agreement, at least in the majority of those cases where the conominal is more informative than the argument index. In other words, gram-indexed verbs can indeed be said to "carry morphological features that originate somewhere else" (cf. Bickel and Nichols 2007), namely the conominal. Examples of gram-indexing system can be found in German or Russian. While Haspelmath's three types of person indexing systems are defined solely in terms of conominal, it is also possible (and potentially revealing) to investigate them in terms of other parameters. For this we introduce some of the parameters that have figured in Corbett's notion of canonical agreement.

[8] Graded approaches to agreement are widespread in the typological literature, e.g. Mithun (2003), Siewierska (1999) and Nichols (1986).

[9] cf. Witzlack-Makarevich and Giorgo 2013 for critical assessment of determining clause-internal vs. clause-external position of relevant expressions

3.2.1 Corbett's concept of 'canonical agreement'

Corbett (2003) develops an approach to agreement which assumes the existence of a controller and a target, and attempts to characterize the nature of the relationship between the two along a number of logically independent parameters. On this approach, different agreement systems can be classified as being closer to, or further away from, the pole of maximal, or canonical agreement. Corbett's framework is not restricted to person indexing between verbs and their arguments, but applies to a broader range of agreement relationships. Nevertheless, they coincide to some extent with Haspelmath's typology, and introduce further distinctions which may be relevant for analysts working on person indexing. We present a selection of these parameters below.

Multirepresentation and obligatoriness. If an agreement configuration requires both controller and target to be overtly present, then it is closer to the pole of canonical agreement. Corbett refers to this parameter as multirepresentation. For person indexing, this basically coincides with Haspelmath's gram-indexing, because it requires the presence of an overt argument (e.g. NP or free pronoun) as well as the bound person index. A second feature of canonical agreement is 'obligatoriness' in that the target is realized by a particular syntactic configuration, for example a particular tense form of a verb, or a clause type—regardless of any discourse factors. Although obligatoriness and multirepresentation generally go hand in hand, we nevertheless consider them distinct: An obligatory agreement marker is still an obligatory agreement marker, regardless of whether it occurs in the presence of a free conominal or not. For example, in German, person marking on finite clauses is obligatory, but in coordinated clauses, a free pronoun is not obligatory in the second conjunct. As noted above in our recommendations for annotators, obligatoriness is for us the more important factor, and will determine whether a bound index is to be annotated or not.

Morphological boundedness. A further criterion for canonical agreement is the 'boundedness' of argument indexes: the person forms involved in canonical agreement (and the gram-index systems in German and Russian) are affixes, i.e. phonologically bound formatives that (a) are inseparable from their host; (b) are restricted to a single category of host (i.e. verbs in these cases); (c) behave in terms of morphophonological processes such as vowel harmony like other inflectional morphemes (e.g. case markers); (d) exhibit allomorphy determined by other inflectional dimensions of the predicate (that is, there are often distinct paradigms for person/number agreement depending on, for example, the tense of the verb); (e) undergo phonological fusion with their host.

Referentiality and descriptive content. Moreover, canonical agreement shares characteristics in terms of 'referentiality' and 'descriptive content' with gram-indexing systems in German and Russian: canonical agreement is not capable of signalling discourse information such as contrastive focus, definiteness etc. Furthermore, agreement need not have any obvious referential function, as when third person singular agreement in German is used as the default agreement with, e.g. weather verbs, impersonal expressions of obligation, or passivized intransitives lacking a referential subject (note that Russian differs

from German in this regard in not allowing free (conominal) subject pronoun in these contexts). The deployment of the free pronoun in the same clause is a matter of pragmatics: focus, contrastive negation, or other factors. Its presence or absence in a given clause does not alter the status of an obligatory agreement affix.

With canonical agreement, the target carries only a minimum of descriptive information. This generally also holds for gram-indexing. For example, third person singular agreement on German verbs has a single form, regardless of gender of the controller, while third person singular free pronouns distinguish three genders. Another distinction, to our knowledge not mentioned in the literature, is the expression of WH-questions, as in English *what did you see / take / find?* etc. In order to express the question, the third person P argument in sentences of this sort needs to be overtly expressed by an interrogative or indefinite pronoun or NP. We are not, however, aware of any language where a person form resembling canonical agreement / gram-indexing may carry the distinction between simple statement and content-question; neither are we aware of any such person form that would, by itself, distinguish between a referential P in a declarative clause, and the questioned P in an interrogative clause.

Argument roles Canonical agreement systems are typically restricted to a single 'index set' (in terms of Haspelmath 2013), as is the case with the gram-indexing systems in German and Russian. And as in these two languages, the syntactic function most commonly instantiated by canonical agreement is S, usually combining with the role of A. But canonical agreement with S and P is also possible, as in Hinuq (Nakh-Daghestanian, Daghestan, Forker 2010). In this language, verbs in "all simple clause types" agree with the argument in the Absolutive case, generally either S or P (Forker 2010:420)

To sum up, canonical agreement is essentially the exponent of a purely syntactic process, a more or less mechanical replication of features of an obligatorily present controller within the same clause. In the realm of person marking, that means that the target is typically realized on the verb, and is oblivious to pragmatics. For precisely this reason, canonical agreement will not normally need to be glossed in GRAID, because its presence is independent of discourse considerations.

It is, however, important to note that in actual language usage even the most 'grammatical' or 'canonical' systems, as in German or Russian, display considerable variation between an ideal gram-system and a cross-system. We can phrase this differently in terms of Bickel's 2003 notion of Referential Density (RD): no language, not even German, English, or Russian, has a RD of 100%. Characterising different gram-indexing systems along the lines of *actual* conomination is one of the empirical questions that GRAID is intended to tackle.

3.2.2 Haspelmath's 'Cross-indexes'

In cross-index systems, conominals may co-occur with the argument index in the same narrow clause, but in contrast to gram-index system, they are not obligatory. The possible absence of a conominal is what has lead to a lot of confusion about the notion of 'agreement' in these systems and in linguistic theory in general, as summarised by Haspelmath (2013). These are also the

systems most challenging for GRAID annotations, and in fact, are probably typologically the most widespread system in the languages of the world. Cross-index systems systems deviate from Corbett's canonical agreement because they do not **require** multirepresentation, but they also tend to diverge from canonical agreement on other parameters as well. Thus, within the class of cross-index systems we find a broad range of phenomena that will have to be handled in GRAID.

Multirepresentation and obligatoriness. Like gram-index systems, cross-index systems allow for conominals, but also tolerate absence of the conominal. What makes these systems particularly interesting for GRAID annotations is that there appears to be considerable cross-language variation here in the extent to which conominals are present, and precisely this kind of variation can best be captured in a corpus-based quantitative approach. In fact, there is probably no hard-and-fast distinction between gram-index and cross-index systems, but rather a continuum of increasing levels of conomination, with languages like German representing simply the higher end of the scale.

Morphological boundedness. Cross-indexes often do involve affixal person forms, bringing such systems closer to canonical agreement. But they often also involve clitic person forms. There is a great deal of cross-linguistic variation in this respect, and annotators should be wary of across-the-board solutions.

Referentiality and descriptive content. Where cross-indexes do co-occur with conominals, their functional load in terms of referentiality is naturally fairly low, similar to that of gram-indexes generally. Nevertheless, they are by definition often the only means overt expression of an argument function, and in such cases could be considered to bear greater referential function. It is this difference in referential function that lead Bresnan and McHombo (1987) (among others) to the (in our view) aberrant analysis that the two contexts (+conomination vs. –conomination) in fact involve two different types of person forms, one being an agreement marker, and the other a kind of (bound) pronoun. We do not endorse this view (see Haspelmath (2013) for critical assessment); instead, we annotate the bound person forms (the cross-index) in the same manner, regardless of the presence or absence of a conominal. Where both are present, we consider them as simply two means of expression for the argument role in question, each with different formal properties. As opposed to free pronouns, which may be used for contrastive purposes, and may also carry information on discourse status (definite vs. indefinite), cross-indexes are not marked for such distinctions. In fact, as mentioned above, it is the suspension of cross-indexes that serves the endcoding of information status and information structural role; and it is the additional use of conominal free pronouns (or pronominal phrases, as in Vera'a and other Oceanic languages) that serves the marking of pragmatic contrast. Like gram-indexes, cross-indexes are to our knowledge also not capable of marking distinctions in illocutionary force.

Languages may vary widely as to what kind of descriptive information (like gender/sexus, number/multitude, etc.) may be coded in cross-indexes as compared to head nouns and other constituents of the corresponding NP (Haspelmath 2013, Siewierska 1999, and others). Often, however, they contrast with

free pronouns in other languages, which are usually capable of finer distinctions than cross-indexes. As with referentiality, possible mis-balances in this regard will not alter the status of either the index or the conominal.

Argument roles. Cross-indexes are not confined to a single set indexing a single argument role (or combination of these, as the frequent S+A subject indexing pattern). Instead, they may cross-index up to three roles (Yimas, Manam), and most typically index at least two (typically S/A and P; cf. Siewierska and Bakker 2013, 1999). As Haspelmath (2013) outlines, languages may have either a single set of cross-indexes arranged differently within the verb form to mark different argument roles (cf. Bantu languages), or different sets for different roles (Sakapultek, cf. below). Rare cases of P-only indexing that come close to canonical agreement are probably also cross-indexes, for instance Savosavo (Papuan,Wegener 2008).

3.2.3 Haspelmath's 'Pro-indexes'

Pro-indexes do not allow for conominals. Hence, these person forms never enter any coreference relation with a conominal in the same clause that could be labelled 'agreement'. In fact, these indexes resemble free personal pronouns in most regards.

Multirepresentation and obligatoriness. Multipresentation is by definition precluded for pro-indexes. From this it also follows that they are not obligatory (their presence in a clause is mediated by the presence of another element). Even when a co-referential free expression is absent, they need not always occur, and the extent of their occurrence is an open question. Certainly there exist languages where the use of pro-indexes is conditioned or suspended, as is the case with some cross-index systems, Vera'a being an example in place. Pro-indexes are in principle in free variation with free pronouns for pronominal reference, for instance in Anejom (cf. Lynch 2000), but also in Vera'a (where the free person form is a pronominal NP). Nevertheless, pro-indexes appear to be the (pragmatically) unmarked choice of pronominal reference in these languages.

Morphological boundedness. Pro-indexes may be affixes or clitics, like cross-indexes. But in some languages, like Vera'a, they seem to occupy a position between free and bound form: Vera'a uses a single set of ("free") person forms for the functions S, A, P and complements of prepositions. The form for P arguments is a pro-index according to the criteria above. It is not free in the sense that it is reduced to a form that cannot be used on its own in an utterance by itself, and that does not carry its own stress. Instead, it is incorporated into the verb complex, and receives stress according to phrasal accent placement or not. It is on the other hand not morphologically bound. Thus, it seems that pro-indexes can in some cases come quite close to free personal pronouns in terms of their morphological and phonological (non-)boundedness.

Referentiality and descriptive content. Pro-indexes show referential characteristics similar to those of cross-indexes. As just outlined, pro-indexes in some languages occur in free variation with free pronominal expressions on

clause level, and the latter type of expression would be used in for instance contrastive or otherwise pragmatically marked contexts. In terms of descriptive information, pro-indexes are generally not different from free personal pronouns.

Argument roles In principle, it seems, pro-indexes are similar to cross-indexes in allowing for more than one argument role to be triggered. In practice, however, it seems that where a language has indexing for more than one (set of) role(s), one will involve cross-indexes and another pro-indexes. This is the case for Central Kurdish past transitive clauses, where A is cross-indexed, but P and recipients are expressed through pro-indexes. We are not aware of a language where, for instance, both S/A and other arguments show pro-indexing.

4 Specific issues of analysis

In this section we broach various issues of analysis that do not fit readily with preceding sections, and make some suggestions for resolving them.

4.1 Identifying clause units

When glossing a text with GRAID, the annotator has to have an idea about how many predicates / clauses, arguments and argument positions s/he assumes to be present in a particular construction and how to apply the glosses available accordingly. Though in principle these analytical decisions must be left to the expert for a given language, we present some general conventions concerning a number of problematic cases. To some extent we adopt Bickel's (2003:721–722) conventions for counting clauses and arguments. However, some modifications will be discussed briefly in what follows.

The basic unit for GRAID is the clause unit, consisting of a predicate and its arguments. Dependent, in particular embedded, clauses pose a difficulty for GRAID annotations, because in terms of their external syntactic function, they are comparable to nominal arguments, yet they have their own internal syntax, including some form of predicate. As outlined in Section 2.6 above, we gloss them as clause units, but add a symbol for the clause type to the clause-boundary symbol <#>, and—optionally—an indication of its external function (e.g. <#cc:p...> for a complement clause bearing P function in the matrix clause).

Predicates consisting of several distinct lexemes, for example serial verb constructions, certain modal expressions, or light verb constructions, can also be problematic. Investigators must make language-specific decisions on whether to count these constructions as a single predicate, in which case they are simply glossed <v:pred> under the head word (verb). Sub-constituents of the predicate (auxiliary particles, the nominal components of light verb constructions etc.) can be either left unglossed, or glossed <other>, or glossed with the left/right glosses discussed in Section 2.9. The alternative, particularly for modal verbs+full verb constructions, or serial verbs, is to count individual verb forms as multiple individual predicates, in which case the annotator would be obliged to set up more than one clause unit. The decisions on this are notoriously fraught. The general spirit of GRAID annotations suggests that when

multi-lexeme predicates behave on most distributional properties like simplex predicates, then they should be glossed as simplex predicates.

Where the repetition of uninflected—and in certain instances also inflected—verb forms or verbal phrases has the function of expressing the duration of an event, or its intensity, and does not impact on the argument structure or help to structure the discourse, we treat the entire series as one predicate. Tail-head linkages, common for example in Oceanic, on the other hand, are treated as constituting separate (non-finite) clauses. Though they often take up an event that has already been mentioned in a preceding clause, these constructions frequently serve to shape the course of the narrative, so we would prefer to gloss them as separate clause units. This remains, however, a topic for future research.

4.2 Reflexive and reciprocal constructions

Languages differ considerably in the means that they express reflexive states of affairs. Some may use a pronoun, which may be identical in form to the corresponding non-reflexive person form (e.g. German *mich* 'me, myself') while others have dedicated reflexive pronouns, as in Engl. *myself*. Some use an affix on the verb, as in Turkish *tara-n-* 'comb oneself', where the *-n* suffix indicates reflexivity. Other languages may leave reflexivity unmarked in many contexts (as in English *he shaved*, where the default reading implies 'himself'), or languages may combine these strategies in various ways (German has in fact a distinct 3rd person reflexive pronoun). These strategies may also then be extended for use with the expression of non-reflexive states-of-affairs and other types of predicates. For example, German uses reflexive pronouns with the verbs for 'remember' (*sich erinnern*) and 'be happy' (*sich freuen*), while many other languages do not treat these predicates as reflexive. Reflexivity raises the following questions for GRAID annotators:

1. Is a reflexive verb such as Turkish *tara-n-* 'comb oneself' to be considered transitive or intransitive? This will affect whether the subject is coded as S, or as A.

2. Is an overt reflexive pronoun to be considered a pronominal argument or not?

Given the many variables involved, it is impossible to propose across-the-board solutions for all languages. As a general recommendation, it seems reasonable to count a reflexive pronoun as an argument when it is used with a transitive verb that is most commonly not reflexive, as in for example *he saw himself in the mirror*. In this case, the pronoun could be glossed <refl.1:P>.

For verbs with a lexically reflexive meaning, or at least a strong cultural implicature for a reflexive reading (for example verbs of grooming, such as *comb, wash, shave*), we would generally advise treating them as intransitive verbs, unless accompanied by an overt reflexive pronoun. In the latter case, for instance with German *Er wäscht sich*, we recommend glossing the reflexive pronoun as <refl.h:p>. For languages which code reflexivity through verbal affixes, with no additional reflexive pronoun provided, it may be appropriate to consider the affix as a valency-reducing device, yielding an intransitive verb whose subject will be coded as S.

In principle, similar issues arise with reciprocity where languages either use reflexive or plain person forms, or both, or none (e.g. Cantonese). We extend the above recommendations to reciprocal constructions, and reciprocal pronouns can also be glossed using the symbol <refl>; there is no dedicated tag for reciprocal pronouns in GRAID.

4.3 Argument positions with non-finite predicates

As indicated above, a number of non-finite constructions, for instance participial, infinitival or converb constructions, categorically block the overt expression of one (most often the 'highest-ranking') argument (generally subject), so that overt realisation is not an option here. As explained above, we adopt Bickel (2003) position that in such cases of systematic blocking of an argument, no argument position is available and and we would therefore gloss 'zero'. Consider the following examples:

(46) English
  a. *I promised my mother [to sell the motorbike]*
  b. *I promised my mother [that I would sell the motorbike]*

According to the rule just given, the two semantically very similar bracketed clauses in (46a) and (46b) will have to be glossed differently: in (46a) there will be just one argument, <np:p>. The predicate would be glossed with <vother:pred>, which indicates that it is a predicate, but one which lacks the normal range of possibilities for assigning argument roles. In (46b), we would have a finite predicate <v:pred> and two arguments, <pro.1:a> and <np:p>).

4.4 Complement clauses

Complement clauses embody a paradox: On the one hand, they exhibit a similar distribution to certain types of NP arguments, i.e. fill argument positions. Thus they have an external function with regard to the matrix predicate. On the other hand, they have their own internal predicate-argument structure. This raises certain problems for annotators, which we consider in this section.

The lefthand border of a complement clause can be indicated in GRAID using the symbol <cc> 'complement clause, which is written immediately following the clause boundary symbol <#>, i.e. <#cc> (cf. Section 2.6 for details). In order to note the external function of the complement clause with regard to its matrix clause, the syntactic function symbols <s>, <a>, <p>, etc. are used in the same way as with argument NPs, for example: <#cc:p>. Thus, the English complex sentences in (47a) could be glossed as follows in GRAID:[10]

(47) English
  a. *That Shawn came to the party surprised me.*
  #cc:a other np.h:s v:pred adp np:g # v:pred pro.1:p

[10] The example (47a) could also be with an expletive pronoun, as in *It surprised me that Shawn came to the party*, which would indeed be more natural in some contexts. The choice between the two constructions will quite certainly be influenced by considerations of information structure. Glossing of such expletive elements can be achieved in a number of ways, the simplest being for this example <other:A>. The complement clause would then have either no overt function gloss (just cc), or <cc:other>.

b. *Irv believes that Harriet is a secret agent.*
np.h:a v:pred #cc:p other np.h:s cop np:pred

Note, however, that the solutions suggested depend on certain assumptions, which themselves are somewhat controversial. For example, English object clauses introduced by *that* display somewhat different syntactic properties from those of an NP argument in P-function. For instance, some transitive verbs do not allow *that*-clauses as P arguments, although they nevertheless occur as the S argument of a passive clause with the very same verb (cf. Bresnan 2001:17):

(48) English
a. **This theory captures [that languages are learnable].*
b. *[That languages are learnable] is captured by this theory.*

On the other hand, for a number of speakers of English, *that*-clauses cannot be promoted to subjects under passivization of a cognition verb like *believe*:

(49) ?? *That Harriet is a secret agent is believed by Irv.*

Thus there is some doubt as to whether such complement clauses really do qualify as P-arguments. If the annotator wishes to record that a particular construction is a complement clause, but considers the syntactic function of the clause to be uncertain, it can be glossed with <#cc:other>. Note, however, that if this solution is chosen, the clause would lack a P-argument and would best be considered intransitive. Thus the subject would be glossed as S rather than A. Applying this option would yield the following gloss for (47b):

(47b') *Irv believes that Harriet is a secret agent.*
np.h:s v:pred #cc:other other np.h:s cop np:pred

The same problems arise with complementation by direct speech. Direct speech diverts in some sense even more drastically from NP complementation, as discussed in 2.6.4 above. Therefore, our position here is that direct speech probably never qualifies as a regular P argument - even though it may occupy the same functional slot as a NP complement would. Only NP complements of verbs of saying are considered P arguments in transitive clauses, and glossing would thus look as follows in any language with euqivalent structures:

(50) a. *Irv says ' I hate our neighbours!'*
np.h:s_ds v:pred ##ds pro.1:a v:pred ln np.h:p

Irv often says things that upset her.'
np.h:s other v:pred np:p #rc other 0:a v:pred pro.h:p

Thus, NPs denoting an utterer of direct speech are considered S arguments, but in order to keep track of their possibly special status they take the additinal information gloss <_ds>, the gloss <:s_ds > meaning 'S argument of a verb of saying with direct speech complement'. This practice then allows to later differentiate between 'regular' S arguments and those that pattern with A arguments in the sense outlined here. Where the same verb takes a content-denoting NP complement instead of direct speech, the construction may still qualify as transitive, containing a A and a P argument, as is the case in the b. example here.

Further problems arise with the centre-embedding of complement clauses, solutions to which are discussed above in Section 4.1.

4.4.1 Syntactically ambiguous arguments: raising and related issues

Some complex clauses involve arguments which are syntactically 'ambiguous', that is, which can be considered to belong to different clauses, depending on the analysis chosen. For English, these issues have been extensively discussed under the label of 'subject-to-object raising'. Consider the following examples from Noonan (2007:79); the GRAID-glossing of the a-example further illustrates the two possibilities discussed in the preceding example:

(51) English

a. *Irv believes [Harriet is a secret agent].*
   \# np.h:a v:pred #cc:p np.h:s cop ln ln np.h:pred
   or # np.h:s v:pred #cc:other np.h:s cop ln ln np.h:pred

b. *Irv believes Harriet [to be a secret agent].*
   \# np.h:a v:pred np.h:p #cc:other ln vother:pred ln ln np.h:other

In (51a), *Harriet* is fairly clearly an argument of the complement clause (it controls agreement on the verb, and if pronominalized, it takes the subject form of the pronoun *she*). In (51b), on the other hand, there is good evidence that *Harriet* is in fact the object of *believes*: under pronominalization, *Harriet* takes on the object form *her*, and *Harriet* can also be promoted to subject under passivization:

(52) English

a. *Irv believes her to be a secret agent.*
b. *Harriet is believed to be a secret agent.*

Thus despite the underlying semantics, this fairly straightforward syntactic evidence does indeed suggest that *Harriet* in (51b) should be glossed as an object to *believes*.

Another problem of clausal loyalty is the so-called 'raising construction' in English. Under raising as it is commonly understood, a semantic argument of the predicate of a complement clause is not realized syntactically within the complement clause itself, but as a syntactic argument of the matrix clause predicate, although it does not bear a thematic relation to the latter. The clearest example of raising—and quite possibly the only pure instance of raising in English—involves the verb *seem*:

(53) English

a. *It seems [that Harriet is a secret agent].*
b. *Harriet seems [to be a secret agent].*

In (53a) *Harriet* is an argument of the complement clause following the complementizer *that*. In (53b), *Harriet* is apparently 'raised' from the subject position of the complement clause to the subject position of the matrix clause. This type of raising construction has played a pivotal role in the development of generative syntactic theory across the last half a century, yet our impression is that outside of Standard Average European, it is actually quite rare, and hence unlikely to be more than a marginal phenomenon for most annotators.

More generally, in keeping with the spirit of GRAID, where surface syntactic configurations are taken at face value, we would consider the argument *Harriet* in (53b) to be the S of *seems*, while the complement clause would be glossed with <#cc:other>. Whether a zero-argument is then included in the non-finite complement clause will depend on which decision the annotator has made for dealing with non-finite predicates (cf. the discussion in Section 4.3 in connection with (46) above).

In sum, it is quite often the case that verbs of perception or cognition take arguments that can be construed both as objects of the main verb, or subjects of a subordinate verb. Our recommendation is that the surface morphosyntax of such arguments be given the highest priority in deciding how to gloss them.

References

Andrews, Avery. 2007. The major functions of the nouun phrase. In *Language typology and syntactic description*, ed. Timothy Shopen, 132–223. Cambridge: Cambridge University Press.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents.*. London: Routledge.

Ariel, Mira. 2000. The development of person agreement markers: From pronouns to higher accessibility markers. In *Usage-based models of language*, ed. Michael Barlow and Suzanne Kemmer, 197–260. Stanford: Centre for the Study of Language and Information.

Bauer, Winifred. 1993. *Maori*. London and New York.

Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79(4):708–736.

Bickel, Balthasar. 2011. Grammatical relations typology. In *The oxford handbook of linguistic typology*, ed. Jae Jung Song, 399–444. Oxford: Oxford University Press.

Bickel, Balthasar, and Johanna Nichols. 2007. Inflectional morphology. In *Language typology and syntactic description*, ed. Timothy Shopen, 169–240. Cambridge: Cambridge University Press.

Bickel, Balthasar, and Johanna Nichols. 2009. Case marking and alignment. In *The oxford handbook of case*, ed. Andrej Malchukov and Timothy Shopen, 304–321. Oxford: Oxford University Press.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Oxford, Cambridge (MA): Blackwell.

Bresnan, Joan, and Sam McHombo. 1987. Topic, pronoun, and agreement in chichewa. *Language* 63 (4):741–782. URL `http://www.jstor.org/stable/415717`.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In *Subject and topic*, ed. Charles N. Li, 25–55. New York: Academic Press.

Chafe, Wallace L. 1994. *Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Comrie, Bernard. 1989. *Language universals and linguistic typology. second edition*. Chicago: The University of Chicago Press.

Comrie, Bernard. 2001. Different views of language typology. In *Language*

typology and language universals, an international handbook, ed. Martin and Haspelmath, 25–39. Berlin, New York: Mouton de Gruyter.

Comrie, Bernard, Martin Haspelmath, and Balthasar Bickel. 2008. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Corbett, Greville C. 2003. Agreement: The range of the phenomenon and the principles of the Surrey Database of Agreement. In *Agreement: A typological perspective (special number of Transactions of the Philological Society 101, no. 2)*, ed. Dunstan Brown, Greville C. Corbett, and Carole Tiberius, 155–202,.

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Dixon, R.M.W. 2010. *Basic linguistic theory, volume 2: Grammatical topics*. Oxford: Oxford University Press.

Dixon, Robert M. W., and Alexandra Y. Aikhenvald. 2002. Word: A typological framework. In *Word: A cross-linguistic typology*, ed. Robert M. W. Dixon and Alexandra Y. Aikhenvald, 1–41. Cambridge: Cambridge University Press.

Donohue, Mark. 2008. Semantic alignment: what's what and what's not. In *The typology of semantic alignment*, ed. Mark Donohue and Sören Wichman, 24–75. Oxford: Oxford University Press.

Dryer, Matthew S. 1986. Primary objects, secondary objects, and antidative. *Language* 62:808–845.

Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63(4):805–855.

Du Bois, John W., Lorraine E. Kumpf, and William J. Ashby, ed. 2003. *Preferred argument structure: Grammar as architecture for function*. Amsterdam, Philadelphia: John Benjamins.

Evans, Nicholas. 1999. Why argument affixes in polysynthetic languages are not pronouuns: eveidence from bininj gun-wok. *Sprachtypologie und Universalienforschung STUF* 52:255–281.

Farrell, Patrick. 2005. *Grammatical relations*. Oxford: Oxford University Press.

Foley, William A. 1991. *The Yimas language of New Guinea*. Stanford: Stanford University Press.

Forker, Diana. 2010. A grammar of Hinuq. Ph.d. dissertation, Universität Leipzig, Philologische Fakultät der Universität Leipzig.

Haig, Geoffrey. 2008. *Alignment change in Iranian languages. a Construction Grammar Approach*. Berlin, New York: Mouton de Gruyter.

Haig, Geoffrey. 2009. On the proposed correlation between discourse pro-drop and fusional pronominal case: Evidence from Iranian. Presentation at the *Third International Conference on Iranian Linguistics*, Paris Sorbonne 3 (September 2009).

Haig, Geoffrey, Ludwig Paul, and Philip Kreyenbroek, ed. 2011a. *Documentation of Gorani, an endangered language of West Iran*. Nijmegen: DoBeS. `http://www.mpi.nl/DOBES/projects/gorani`.

Haig, Geoffrey, and Stefan Schnell. 2009. From text to typology: Towards implementing quantitative typology on corpora from endangered languages. In *Proceedings of the language documentation and linguistic theory 2 conference*, ed. Peter K. Austin and Oliver Bond, 117–122. London: School of Oriental and African Studies. `http://www.hrelp.org/publications/ldlt2/papers/ldlt2_12.pdf`.

Haig, Geoffrey, Stefan Schnell, and Claudia Wegener. 2011b. Comparing corpora from endangered language projects: Explorations in typology with original

texts. Ms., available on demand from the authors.
Haspelmath, Martin. 2011. On S, A, P, T and R as comparative concepts for alignment typology. *Linguistic Typology* 15.
Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntactic status of bound person forms. In *Languages across boundaries: Studies in memory of anne siewierska*, ed. Dik Bakker and Haspelmath Martin, 197–226. Berlin, New York: Mouton de Gruyter.
Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195.
Himmelmann, Nikolaus P. 2006. Language documentation: What is it, and what is it good for. In *Essentials of language documentation*, ed. Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, Trends in Linguistics: Studies and Monographs 178, 1–30. Berlin, New York: Mouton de Gruyter.
Jelinek, Eloise. 1984. Empty categories and non-configurational languages. *Natural Language and Linguistic Theory* 2:39–76.
Jukes, Anthony. 2005. Makassar. In *The Austronesian languages of Asia and Madagascar*, ed. Nikolaus P. Himmelmann and Alexander Adelaar, 662. London: Routledge.
Kibrik, Andrej Aleksandrovič. 2011. *Reference in discourse*. Oxford: Oxford University Press.
Koptjevskaja-Tamm, Maria. 1993. *Nominalizations*. London, New York: Routledge.
Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus and the mental representation of discourse referents*. Cambridge Studies in Linguistics 71. Cambridge: Cambridge University Press.
Lynch, John. 2000. *A grammar of anejom*. Canberra: Pacific Linguistics, Research Schools of Pacific and Asian Studies.
Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.
Malchukov, Andrej, Martin Haspelmath, and Bernard Comrie. 2010. Ditransitive constructions: a typological overview. In *Studies in ditransitive constructions: a comparative handbook*, ed. Andrej Malchukov, Martin Haspelmath, and Bernard Comrie, 1–60. Berlin, New York: Mouton de Gruyter.
Margetts, Anna. 2004. From implicature to construction: Emergence of a benefactive construction in oceanic. *Oceanic Linguistics* 43 (2):445–468.
Margetts, Anna. 2007. Three-participant events in the languages of the world: Towards a crosslinguistic typology. *Linguistics* 45(3):393–451.
Mithun, Marianne. 1984. The evolution of noun incorporation. *Language* 60:847–894.
Mithun, Marianne. 2003. Pronouns and agreement: the information status of pronominal affixes. *Transactions of the Philological Society* 101(2):235–278.
Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1):56–119.
Noonan, Michael. 2003. A cross-linguistic investigation of referential density. Online publication, available at `http://archiv.ub.uni-heidelberg.de/savifadok/volltexte/2008/190/`.
Noonan, Michael. 2007. Complementation. In *Language typology and syntactic description*, ed. Timothy Shopen, volume 3: Complex Constructions, 51–150. Cambridge: Cambridge University Press, second edition.
Payne, Thomas E. 1992. *The twins stories: Participant coding in Yagua narra-*

tive. Berkeley: University of California Press.
Peterson, John. 2011. *A grammar of Kharia: A South Munda language*. Brill's Studies in South and Southwest Asian Languages (BSSAL), 1. Leiden: Brill.
Schiering, René, Balthasar Bickel, and Kristine Hildebrandt. 2010. The prosodic word is not universal but emergent. *Journal of Linguistics* 46:657–709.
Schnell, Stefan, and Geoffrey Haig. In print. Assessing the relationship between object topicalisation and the grammaticalisation of object agreement. In *Selected Papers from the 44th Conference of the Australian Linguistic Society, 2013*.
Schultze-Berndt, Eva. 2006. Linguistic annotation. In *Essentials of language documentation*, ed. Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, Trends in Linguistics: Studies and Monographs 178, 213–251. Berlin, New York: Mouton de Gruyter.
Seifart, Frank, Roland Meyer, Taras Zakharko, Balthasar Bickel, Swintha Danielson, and Alena Witzlack-Makarevich. 2010. Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis. Presentation at the workshop *Advances in Documentary Linguistics*, MPI Nijmegen, 14–15 October 2010.
Siewierska, Anna. 1999. From anaphoric pronoun to grammatical agreement marker: why objects don't make it. *Folia linguistica* 33(2):225–251.
Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.
Siewierska, Anna, and Dik Bakker. 2013. Suppletion in person forms: The role of iconicity and frequency. In *Languages across boundaries: Studies in memory of anne siewierska*, ed. Dik Bakker and Haspelmath Martin, 347–396. Berlin, New York: Mouton de Gruyter.
Stoll, Sabine, and Balthasar Bickel. 2009. How deep are differences in referential density? In *Crosslinguistic approaches to the psychology of language: Research in the tradition of dan isaac slobin*, ed. Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Şeyda Özçalişkan, Psychology Press Festschrift Series, 543–555. London: Psychology Press.
Van Valin, Robert D., and Randy J. LaPolla. 1997. *Syntax*. Cambridge: Cambridge University Press.
Van Valin, Robert D., Jr. 2005. *Exploring the syntax-semantics interface*. Cambridge: Cambridge University Press.
Wegener, Claudia. 2008. *A grammar of Savosavo, a Papuan language of the Solomon Islands*. MPI Series in Psycholinguistics 51. Nijmegen: MPI for Psycholinguistics.
Witzlack-Makarevich, Alena, and Iemmolo Giorgo. 2013. When is there agreement? typologizing suspension restrictions on agreement. Presentation at Association of Linguistic Typology Conference.

5 Alphabetical list of GRAID symbols

##	boundary of main, syntactically independent clause
#	boundary of all other clauses
\	separates two person values of a bound formative that expresses a particular combination of two arguments, cf. Section 3.1
0	'zero': argument position not filled by an overt referring expression
1	argument with 1st person referent(s)
2	argument with 2nd person referent(s)
a	transitive subject
ac	adverbial clause
adp	adposition
appos	apposition, cf. end of Section 2.4
aux	auxiliary
cc	complement clause
cop	overt copular verb, in combination with some kind of non-verbal predicate complement, cf. Section 2.5.2
d	optional; can be used to distinguish genuine human referents from those with anthropomorphized referent(s), e.g. spirits, mythical figures, capable of speech and self reference.
ds	direct speech
dt	dislocated topic
g	goal argument of a goal-oriented verb of motion, transitive or intransitive, may also extend to Recipient and Addressee, cf. Section 2.4
h	NP has human referent(s), or refers to anthropomorphized referents
l	locative argument of verbs of location
ln	NP-internal subconstituent occurring to the left of NP head
lv	subconstituent of verb complex occurring to the left of verbal head
nc	'not considered' / 'non-classifiable'
ncs	non-canonical subject: An argument that lacks some or all of the morphological properties associated with subjects in the language, but commands most of the syntactic properties associated with subjects in the language concerned
neg	negated
np	lexical NP
obl	oblique argument, excluding goals and locatives
other	other forms / words / functions which are not relevant
p	transitive object
poss	possessor
pred	function gloss for the item that constitutes the predicate of a clause
predex	predicate function in an existential expression
pro	free pronoun in its full form (in contrast to <-pro> or <=pro>)
rc	relative clause
refl	overt reflexive or reciprocal pronoun, cf. Section 4.2

continued on next page

rn	NP-internal subconstituent occurring to the right of NP head
rv	subconstituent of verb complex occurring to the right of verbal head
s	intransitive subject
v	lexical verb as the form element of a predicate
voc	vocative, used for expressions denoting the person to which an utterance is addressed
vother	verbal element, may be used in predicative function, but lacking the normal means for assigning arguments (e.g. certain types of nominalization, imperatives)
w	'weak': Indicates phonologically lighter form of a particular element (e.g. pronoun) that may, under certain conditions, be realized as clitic. Simply precedes regular gloss, e.g. <wpro>