

Annotations using GRAID
(Grammatical Relations and Animacy in Discourse)
Introduction and guidelines for annotators
Version 6.0

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

Stefan Schnell
Kiel University
sschnell@linguistik.uni-kiel.de

September 2011

Contents

1	Introduction	2
1.1	Prerequisites and design of a GRAID data set	3
1.2	Motivation for GRAID glossing	4
1.3	GRAID annotations as basis for quantitative analyses	4
2	Using GRAID annotations in practice	5
2.1	Overview	5
2.2	Forms and inherent properties of referential expressions	8
2.3	Functions of referential expressions	11
2.4	Predicates	15
2.5	Clause boundaries, embedded clauses, and clausal operators	20
2.6	‘Non-classifiable’, and ‘other’	21
3	Specific issues of analysis	22
3.1	Identifying clause units	23
3.2	Agreement morphology vs. bound pronouns	24
3.3	Reflexive and reciprocal pronouns	33
3.4	Argument positions with non-finite predicates	34
3.5	Complement clauses	35
4	Alphabetical list of GRAID symbols	42

1 Introduction

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 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, 2011). 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 & Bickel 2009), and Michael 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 & Schnell (2009) and Haig et al (to appear) for an overview of research topics amenable to GRAID annotations).

GRAID currently uses approx. 30 symbols (cf. §4 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.

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 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. For example, we have completed annotations of more than 2000 clause units in 5 languages within a relatively narrow time frame, with more in progress. 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 2006:243). In the remainder of this section we will outline the prerequisites and motivations for GRAID glossings before we proceed to explaining the annotation system and its application in subsequent

sections.

1.1 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, but they 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 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)

1.2 Motivation for GRAID glossing

For cross-corpus comparisons of grammatical features in discourse, a consistent annotation system is necessary. Unfortunately, the currently most widely-used type of grammatical annotation, that of morpheme-for-morpheme glosses, is not 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 a 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 simply not possible. 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.

In general, we believe that GRAID annotations should abide by the principle of ‘Inclusivity’, as outlined by Corbett (2003:158) in his discussion of the Surrey Database of Agreement: It is preferable to include the maximum amount of information in an annotation that is practicable given the state of one’s knowledge of the language, even though it may not be immediately relevant to the annotator’s own research questions.

1.3 GRAID annotations as basis for quantitative analyses

An important feature of GRAID is that the application of glosses already entails counting 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?
- Regularities of word order of pronominal arguments as compared to NP-arguments
- Whether 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.

2 Using GRAID annotations in practice

In this section, we outline the basic principles of GRAID to give the reader a feel for the system, before providing explications of the full inventory of symbols and more extensive examples, and discussing some problematic issues. A short stretch of annotated text is available for illustrative purposes at:

<http://www.linguistik.uni-kiel.de/GRAID%20glossed%20text%20example.pdf>

2.1 Overview

Throughout this manual we enclose the actual symbols of GRAID in triangular brackets, like this: <<...>. GRAID annotations target major syntactic constituents rather than words. Furthermore, GRAID annotations mainly focus on referential expressions serving as arguments or adjuncts (e.g. A, S, etc.), and - with the exception of pronominal affixes - these are phrases in most cases. This means that many words in actual texts do not need to be glossed at all. Essentially, each item of a GRAID annotation couples an abbreviation of a form (e.g. <pro> ‘full pronoun’), which may additionally have an animacy feature, e.g. ‘human’, with a function. 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 (=full pronoun, human referent, in S function)

In addition, GRAID annotations involve decisions on what elements are part of the same prosodic word, and whether they have affixal or clitic status. This of course presupposes that annotators have reached decisions on what constitutes a prosodic word in their particular language (cf. the discussion on wordhood in e.g. Dixon and Aikhenvald (2002), and the research by Bickel and associates at the University of Leipzig on the typology of the word (cf. Schiering et al. 2010). We follow the conventions of the *Leipzig Glossing Rules* for distinguishing between clitics and affixes: affixes are linked by a dash <->, while clitics are linked by <=>, as illustrated in the following examples (cf. §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:P¹
 ‘s/he carries it’

- b. GERMAN

er hat's gemacht
 he **has=it** do:PTCPL
pro.h:A aux=pro:P² v:pred
 ‘he has done it’

As can be seen, the main verbs in these examples are simply glossed <v:pred> ‘verbal predicate’, likewise consisting of a form and a function tag. As the main target of GRAID annotations is the realization of verbal arguments, glossing of most predicates is comparatively coarse grained. Different types of predicate and glossing conventions are outlined in §2.4 below.

GRAID-glossing presupposes that the annotator has a sound notion of which type of argument belongs to a particular predicate. In many cases, an argument that is

¹ Prosodic word consisting of: pronominal affix with human referent in A-function + verb + pronominal affix, non-human referent, with P-function.

² Prosodic word consisting of: auxiliary verb + clitic pronoun (inanimate) in P function

considered to be required by a particular predicate receives no overt realization in the clause. Argument positions that refer to discourse-retrievable entities, but which lack overt expression through a NP or a pronoun are nevertheless glossed. Their form is rendered in the gloss by <0> (the digit ‘zero’). In (3) below, there is no overt NP in S-function, although the verb subcategorizes for such a participant. Therefore we gloss <0.2:S> (‘deleted pronoun (third person by default, see §2.2), with human referent, in S-function’). In this clause, an overt free pronoun or a full NP subject could be supplied without impairing grammaticality, and without triggering any change to the verb (this is why we do not treat the verbal agreement suffix as “pronominal”; see below). These are the main reasons for considering the argument position to be ‘unfilled’ on the surface (see Section 3.2 for discussion of the status of verbal agreement). The interrogative pronoun ‘why’ is glossed as <other>, the generic symbol for constituents outside of the case frame of the verb, or otherwise not classifiable/not considered relevant (see §2.6 below):

- (3) TURKISH
niçin gel-di-n?
 why come-PST-2S
 # 0.2:S **other** v:pred
 ‘why did you come?’

Certain types of predicate 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 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 §3.1 in connection with the counting of predicates. Clause units are separated by <#>. GRAID recognizes three special cases of clause types: negated clauses, relative clauses, and complement clauses. These are indicated at the left-hand boundary of the clause, for instance <#neg> (cf. 2.5 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 4, a full alphabetical list of all symbols used may be found.

2.2 Forms and inherent 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 rather than adjuncts (see §2.3). The main symbols used for the form of these elements are contained in Table 1.

Table 1: Glosses for the form of referential expressions

np	lexical noun phrase
pro	free pronoun in full form
=pro	‘weak’ clitic pronoun
-pro	pronominal affix, cf. 3.2
0	covert argument / unfilled argument position
refl	reflexive or reciprocal pronoun, cf. Section 3.3
adp	adposition
w	‘weak’, indicates a phonologically lighter form, it precedes the form symbol, e.g. <wpro>
other	used for expressions <ol style="list-style-type: none"> 1. that are not of a type listed above 2. the form of which is not considered relevant

As mentioned above, the hyphen <-> and equal sign <=> indicate affixal and clitic boundary respectively. These are most commonly used in GRAID with pronominal morphemes; however, they may also be optionally used with agreement morphology (cf. §3.2) or with incorporated nouns in polysynthetic languages (cf. (11) below). As for distinguishing clitics from affixes, we follow Bickel & 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.

Where languages have three grades of phonological weight in their pronouns (e.g. French *moi*, *je*, *j’=*), 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> (or <aux>) yielding <wpro, waux> if annotators wish to preserve a three-way distinction.

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):

- (4) GERMANY
- Sie wohn-t in diesem Haus*
 she live-PRES.3S in DEM.NEUT.DAT house
- pro.h:S v:pred **adp np:l**
- ‘She lives in this house’

If an NP or an adpositional phrase is an adjunct rather than an argument, the entire phrase may be simply glossed as <other> without any specifications of form and function (s. below). A further possibility is to analyse its form, but give its function as <other>, yielding for instance <adp np:other>:

- (5) *In winter she lives in that house*
adp np:other pro.h:S v:pred adp np:l

The gloss <other> is also used for any expression that is neither a NP nor a pronoun. 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 uses of <refl> are discussed in §3.3 below.

The symbol <0> is probably the most controversial. The use of this gloss presupposes three conditions. 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 be in a position to make such decisions regarding the valency of predicates. When an argument position is assumed, but not overtly filled, the symbol <0> can be used, depending on the second condition. The second condition is the one briefly mentioned above: in some cases, an argument required by the predicate simply cannot be used in a particular syntactic configuration (e.g. suppression of S or A argument in non-finite clauses). In these cases, we do not recommend using <0>. Rather, we reserve it for those contexts where an overt argument could occur without violating grammaticality, but does not. This is typically the case in narrative texts for

arguments whose reference is considered inferrable from the discourse context. Equi-deletion in English 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 $\langle 0.h:S \rangle$ ‘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 $\langle 0:p \rangle$ 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. In practice, we have found relatively few such contexts. Where annotators are unable to reach a clear decision, they may gloss the entire clause unit with $\langle nc \rangle$ ‘non-classifiable’, which would exclude that particular clause unit from the analysis. This is an option generally available for cases of uncertainty (see below).

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

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

Inherent properties include person, and animacy (human vs. non-human). These are linked to the form glosses using the symbol $\langle . \rangle$, as demonstrated in (1) above. The bare $\langle np \rangle$ or $\langle pro \rangle$ symbol is used where a NP or pronoun is third person and has a non-human referent. With third person human referents, $\langle np.h \rangle$, $\langle pro.h \rangle$ and $\langle 0.h \rangle$ are used. For first and second person referents, $\langle 1 \rangle$ and $\langle 2 \rangle$ are used respectively. As humanness is entailed in reference to speech act participants $\langle h \rangle$ would be redundant in combination with $\langle 1 \rangle$ and $\langle 2 \rangle$, and is therefore not used.

Note that the symbols $\langle 1 \rangle$ and $\langle 2 \rangle$ are also used in the optional glossing of agreement morphology (cf. §3.2). In this function, and here only, the symbol $\langle 3 \rangle$ may also be used to indicate agreement with the third person.

The symbol $\langle d \rangle$ is optionally used with anthropomorphised discourse participants (e.g. $\langle np.d \rangle$). 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.

2.3 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 (we use the terms ‘grammatical relation’ and ‘syntactic function’ largely interchangeably 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).³ 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 cross-linguistic generalizations on the possible shapes of grammars can be formulated (Comrie 1989, Farrell 2005, Andrews 2007, Haspelmath To appear 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
dt	dislocated topic (right or left-dislocated)
obl	oblique argument, excluding goals and locatives
poss	possessor
other	other function

³ 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.

It will be evident that the functions covered by our gloss <g> ‘Goal’ extend beyond the semantic role label GOAL; we discuss the issues surrounding the <g> gloss after example (7) on page 13. 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. Capital S is used for the sole arguments of intransitive verbs, including the subjects of non-verbal or copular clauses. Typically, a S argument takes the form of either A (accusative alignment) or P (ergative alignment), but some special cases exist; see below for some discussion.

In many recent publications the symbols S, A and P (or S, A and O) are used in a broader sense than we use them here. For example, S is sometimes 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), while Dixon (2010:151) extends A and O 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. For S, 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. an nominative or absolutive case, if available).⁴

⁴ 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

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’

A further type of oblique arguments are locatives, <l>. This symbol is used for arguments expressing local roles of static location, and also source. For local goals, we use <g>. Of course languages frequently use the same formal means for coding RECIPIENT and ADDRESSEE as they do for GOAL; in such languages, all three will be glossed <g>. Note that in these cases, RECIPIENTS and ADDRESSEES would receive a gloss for animacy, e.g. <g.h>, so that the distinction between them and purely local goals would still 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>, basically the default glossing for those arguments which are not S, A or P, but which cannot readily be rendered with either <l> or <g>.

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.⁵

different concepts of S, A and P have recently been critically summarized in Haspelmath (To appear), 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.

⁵ 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 argu-

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.

A somewhat different set of problems arises in the case of primary-object languages, where RECIPIENTS are regularly coded as P (cf. also RECIPIENTS in English Primary Object Constructions, or ‘dative-shift constructions’, cf. Malchukov et al. (2010) and Dryer (1986). In this case, as in general in GRAID, **formal morphosyntactic coding properties take precedence over semantics**, and the RECIPIENT will be glossed as <P>, while the THEME would receive either <obl> or <other>. An example from English is the following:

- (8) ENGLISH
*Mum gave **us** sweets.*
 # np.h:A v:pred **pro.1:P** adp np:obl

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. It also behaves like a P argument in that it may be promoted to subject under passivization (*We were given sweets*). In such cases, we take the actual surface form of the sentence at face value, and gloss the function of the pronoun with P accordingly.

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

- (9) ENGLISH
*He expected **me** to leave.*

This example, and related issues, are taken up in sections §3.5 below.

In some languages, verbs may require object complements that are not Ps, but which are also neither GOAL nor LOCATIVE. Examples are the dative complements of German *helfen* ‘help’, or the instrumental-marked complement of Russian *vladet* ‘master, rule’ (the ‘Exceptional Case Marking’ of Generative Grammar). In GRAID, these NPs would receive the function-gloss <obl>. Further examples are verbs expressing 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 arguments that are considered to

ments 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.

be part of the verb’s argument frame, but which differ formally from <P>, and are not <l> or <g>.

The symbol <dt> ‘dislocated topic’ is used for NPs that are either fronted to the clause proper or occur at the right clause boundary that do not have an argument relation in the clause. Colloquial English makes extensive use of such elements, as in (10):

- (10) ENGLISH
Mike, he hates syntax.
 # **np.h:dt** pro.h:A v:pred np:P

Pronouns, nouns or NPs with the function of possessor can be glossed with <poss>. As we are mainly concerned with clause structure here, the glossing of possessors is optional. The main reason for including possessors is that they 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 flow 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*).

For other functions that do not match those mentioned so far there is the option of glossing them with <other>. This gloss is conventionally used, for instance, for NPs in apposition or adjunct NPs (note that in the latter case, the form need not be considered, so that the NP could also simply be glossed <other> rather than <np:other>).

2.4 Predicates

In accordance with the research questions outlined in Haig & Schnell (2009), 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. §2.4.1)
vother	non-canonical verb-form (cf. §2.4.4)
cop	(overt) copular verb (cf. §2.4.2)
aux	auxiliary (cf. §2.4.2)
-aux	suffixal auxiliary
=aux	clitic auxiliary
pred	predicative function

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.

2.4.1 Verbal predicates

For the majority of languages, it is possible on formal grounds to identify a class of lexemes that can be equated with the class of ‘verbs’. Predicates whose semantic core is carried by a member of the verb class in the language concerned are glossed with <v:pred>, as in most examples considered above. It has been claimed, however, that some languages lack verbs as a distinct lexical class (e.g. Kharia (South Munda), Peterson 2011). For such languages, annotators can either decide to still use the gloss <v:pred> as the default gloss for predicates, or they may prefer to use <other:pred>. The latter choice avoids a commitment on the lexical class of the predicate, and is therefore perhaps preferable for languages such as Kharia. The use of the <vother> gloss is discussed in 2.4.4 below.

Whichever gloss is chosen, it will be the gloss for the entirety of elements, whether single words or affixes, that comprise that predicate, i.e. it will also cover TAM-markers, valency-changing devices etc. (cf. for example the ‘verb complex’ in Oceanic). The single gloss may also cover serial verb and light verb constructions. Here investigators must reach language-specific decisions on whether to treat additional elements as part of the same predicate (hence receiving no extra gloss), or a distinct element. 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 predicates contain referential information, as in the case of person agreement (cf. ex. (3) above), or incorporated nouns, they can be linked with <-> or <=>. Exs. (11a) and (11b) contrast a transitive sentence with a free NP in P function and the

corresponding clause with incorporated P:

(11) HUAHTLA (Nahuatl, Uto-Aztecan, Mexico) (Mithun; 1984:860)

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

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

2.4.2 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:

- (12) Identification/Equation:
That woman is the managing director
- (13) Classification:
Bob is a linguist
- (14) Property assignment:
he is very old
- (15) Location:
she is at the market

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, receives the function gloss <pred>, and whichever form gloss is appropriate. For example, in (13) we could gloss <np:pred>. If none of the available form labels are suitable, the predicate can be glossed <other:pred>. Thus the GRAID glossing of (12)–(15) would be:

- (a) np.h:S cop np.h:pred
 (b) np.h:S cop np.h:pred

- (c) pro.h:S cop other:pred
 (d) pro.h:S cop adp other:pred⁶

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 predicate like a regular verb, i.e. <v:pred>. The complement would then receive the function gloss <... :other>:

- (16) *She got hurt*
 pro.h:S v:pred other

In the case of locational relations, it may be preferable to gloss such verbs generally as <v:pred>, and the locative complement as <...:l>; this is a matter for individual judgement.

Existential senses of a copula verb, which often lack a complement (cf. English *there is X*) can either be glossed as <cop>, or <v:pred>, depending on whether they are formally closer to the copular, or to other verbs. Where a specialized type of predicate is used (as in English), it can be optionally glossed with <predex>.

We consider auxiliary verbs, glossed <aux>, 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). The formal properties of copulas and auxiliaries can be marked in the same way as with pronouns, using <->, <=> or <w...>.

2.4.3 Non-verbal predicates

In many languages, predicative expressions such as those illustrated in (12)-(15) above, do not require any overt verbal element. Instead, NPs, APs or PPs expressing the predicate are simply juxtaposed to the subject NP:

- (17) TURKISH
Sevgi mühendis / ev-de
 Sevgi engineer / house-LOC
 np.h:S np:pred / other:pred
 ‘Sevgi is an engineer / at home’

⁶ Form glossing for locational copula complements is a potentially complex area. Given the fact that these items can be considered part of the predicate, rather than an argument, it is probably simpler to gloss them with <other>, as in this example, rather than rendering the full form.

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 predication, and gloss as in (17).

2.4.4 Non-canonical predicates: using the “vother” gloss

Sometimes a particular form, or form class, seems to productively express predication in a language, but it is not a fully-fledged finite verb form. This is typically the case with various types of more or less nominalized verb forms, such as infinitives, participles, or converbs. Such non- (or less) finite verb forms generally admit only a subset of the available TAM distinctions in the language concerned, and, crucially for GRAID, this kind of ‘semi-verbal’ form 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 verb forms 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 verb 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 of such verb forms. 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 v:pred #cc:p vother np:p>.

2.5 Clause boundaries, embedded clauses, and clausal operators

GRAID glosses clause boundaries and also, for embedded clauses, whether they are complement clauses or relative clauses. Also, the syntactic function of an embedded clause in the matrix clause can be added in a similar way as with NPs. The relevant glosses are given in Table 5.

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

#	clause boundary, inserted at left edge, one per clause
rc	relative clause
cc	complement clause
neg	negative polarity

The symbol <#> denotes a clause boundary, and is inserted at the left edge of the clause. Relative and complement clauses combine the boundary gloss with that for relative or complement clauses, giving <#rc> and <#cc>, respectively.

Among clausal operators, we consider only polarity. Negated clauses are glossed with <#neg>, while affirmative clauses receive no special operator. Negated relative and complement clauses are glossed as <#rc.neg> and <#cc.neg>, respectively. The syntactic function of complement clauses is 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 in the same manner that complement clauses do, for instance, <#rc:S>.

It should be noted here that the more detailed glossing of embedded 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 §3.5 below.

2.6 ‘Non-classifiable’, and ‘other’

It is not always possible to reach a principled decision on how to gloss a given object language element. In such cases, GRAID offers two options, shown in Table 6.

Table 6: 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. For these elements annotators have the option of either leaving them unglossed, or using the multi-purpose label <other>. 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. <x:other>, for example with NPs with reference to discourse participants, but with a non-argument function in the clause (e.g. the non-verbal complement of a copular verb, cf. (16) above). The <other> gloss can also be used for forms, for example interrogative pronouns (*who*, *what* etc.). They have a very different discourse and reference function to other pronoun types: 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?*).

The gloss <nc> is intended to be used in a somewhat different way, namely where the annotator(s) is/are not sure how to analyze a particular expression or construction. Thus, the following kinds of difficulties may arise:

- 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 lacks 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 waste 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 and Vera’a 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.

3 Specific issues of analysis

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 potentially problematic cases. To some extent we adopt Bickel’s (2003:721–722) conventions for counting clauses and arguments, implemented in his survey of Referential Density in three languages of the Himalayas. However, some modifications of these will be discussed briefly in what follows.

3.1 Identifying clause units

The basic unit for GRAID is the clause unit, consisting of a predicate and its arguments. Complement 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. We gloss them as clause units, but include the operator <cc> immediately after the clause-boundary symbol <#>. Optionally an indication of external function can be added (e.g. <cc:P ...> for a complement clause that is an object to the main verb). Relative clauses are only provided with an indication of syntactic function when they are headless; attributive relative clauses are simply indicated as such (<rc>). In contrast to Bickel (2003) we recommend glossing all relative clauses - and not only those within predicative NPs in existential clauses - because relative clauses may convey important narrative information.

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

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

Similar problems may arise with center-embedded relative clauses. Glossing in this manner, i.e. more or less word-for-word, would create the wrong impression that there is a clause unit containing only the initial <A>, and it would also leave two <pred>'s in the second string.

In order to alleviate this kind of problem, two solutions are available. The first is to extrapose the GRAID annotation of the embedded clause to the right of the matrix clause. This solution is not particularly elegant, but it preserves the integrity of individual clause units:

(19) # A v:pred #cc:P ... v:pred # ...

The second solution is to enclose the embedded clause in square brackets, e.g. <[...]>, as in (18) above. In general, such center-embeddings do not seem to be very frequent in natural spoken discourse, so the two solutions discussed are unlikely to be deployed very often. Depending on the requirements of the annotator, one of the two solutions should be adopted and applied consistently.

Predicates consisting of several distinct lexemes, for example serial verb constructions, certain modal expressions, or light verb constructions, are also problematic. Investigators must make language-specific decisions on whether to count these items as a

single predicate (in which case they are simply glossed <v:pred>), or whether to count them as more than one predicate, in which case they 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 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 clauses. Though they often take up an event that has already been mentioned in a preceding clause, these constructions are optional and frequently serve to shape the course of the narrative, so we would prefer to gloss them as independent clause units. This remains, however, a topic for future research.

3.2 Agreement morphology vs. bound pronouns

The most contentious issue for annotators is that of deciding which formatives are to be considered pronominal, and which are to be considered agreement in the traditional, narrow sense of the term. Decisions here will have important repercussions for quantitative analysis of, for example, Referential Density, hence it is important to appreciate the issues concerned. The theoretical literature on agreement and related issues provides no ready-made solution; on the contrary, different analyses of one and the same language abound. In what follows, we largely adopt the standpoint of Corbett (2003), who works with a notion of “canonical agreement”, defined in terms of several different criteria. Working outward from this core, different types of related phenomena can be situated on a cline, the farthest end of which is occupied by prosodically independent, anaphoric pronouns.⁷ Before looking at this approach in more detail, it is worth introducing a cover term for the kind of phenomena we are interested in: Argument cross-referencing, or simply cross-referencing. By this we mean any linguistic form that replicates features of a verbal argument such as person and number, and, to varying degrees, gender, definiteness, or honorifics. Pronouns obviously fall under argument cross-referencing, but agreement markers such as third person singular present on English verbs likewise

⁷ Graded approaches to agreement are in fact widespread in the typological literature, e.g. Mithun (2003) Siewierska (1999) Nichols (1986).

does, as do many other types of formatives in the languages of the world. The term “cross-referencing” is essentially equivalent to the broad use of “agreement” in Siewierska (2004), but we prefer to reserve the latter for Corbett’s “canonical agreement”.

3.2.1 Identifying canonical agreement

Canonical agreement is essentially the exponent of a purely syntactic process, a more or less mechanical replication of features of a verbal argument, realized usually on the verb, and 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. A good example of canonical agreement is third person singular marking on the simple present tense of English verbs, which we would not normally consider necessary to gloss in GRAID. However, typologically diverse languages display many different types of argument cross-referencing, and in many cases it is far from straightforward to distinguishing pronominal arguments, which need to be glossed, from canonical agreement. A number of different criteria can be helpful here, which we briefly outline below. It is important to note that the criteria are in principle independent; thus while there are typical clusterings of values across the different criteria, there are also mismatches, and it is precisely these cases which tend to cause difficulties.

Criterion 1: Morphological boundedness and host selection Canonical agreement is expressed through 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 the cases discussed here); (c) behave in terms of morphophonological processes such as vowel harmony like other inflectional morphemes (e.g. case markers); (d) may 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) may undergo phonological fusion with its host.

What this criterion defines is a cline of morphological boundedness, ranging from an inflectional affix over clitics to free pronouns. On this criterion, we would expect canonical agreement to be associated with the phonologically most tightly bound formatives.

Criterion 2: Obligatoriness Canonical agreement is required by a particular syntactic configuration, for example a particular tense form of a verb, or a clause type, requires an agreement affix — regardless of any discourse factors. On our view, this is the most important criterion for distinguishing canonical agreement from other types of argument

cross-referencing. This criterion comes close to Corbett’s criterion of “Multirepresentation”. By this is meant the fact that canonical agreement occurs in the presence of a lexical or pronominal representative of its referent in one and the same clause. 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 pronominal argument or not. 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. There are considerable language-specific differences in the degree to which free pronouns may be required or not (the Null Subject Parameter etc.), and these differences are, to some extent, tangential to the presence or absence of agreement morphology in the language concerned (see Bickel (2003) on this point). These are precisely the kinds of differences that GRAID attempts to capture.

Criterion 3: Referentiality and descriptive content Canonical agreement is usually not available for signalling discourse information such as contrastive focus, definiteness etc. Furthermore, canonical agreement may not even have any obvious referential function, as when third person singular agreement is used as the default agreement with, e.g. weather verbs, impersonal expressions of obligation, or passivized intransitives lacking a referential subject. Free pronouns, on the other hand, may be used for contrastive purposes, and may also carry information on discourse status (definite vs. indefinite). Furthermore, they generally encode a range of semantic distinctions, e.g. gender or number distinctions, that is at least equal to, and usually higher than, the corresponding canonical agreement. For example, third person singular agreement on German verbs has a single form, regardless of gender of the controller, while third person singular 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 in sentences of this sort needs to be overtly expressed by some kind of pronoun, either a specialized WH-word, or an indefinite pronoun. Verbal agreement morphology does not, however, to our knowledge carry the distinction between simple statement and content-question; we are not aware of any kind of agreement affix that would, by itself, distinguish between a referential P in a declarative clause, and the questioned P in an interrogative clause.

Criterion 4: Case roles Canonical agreement, where it occurs, is typically restricted to cross-referencing a single case role. The syntactic function most commonly instantiated by canonical agreement is S. In many languages, the same marker will often indicate agreement with A, as in much of Indo-European. 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). In other languages, e.g. Savosavo (Papuan, Wegener 2008), only P arguments are cross-referenced on the verb. Free pronouns, on the other hand, are generally available for the full range of case distinctions made in the language concerned.

3.2.2 Recommendations for annotators

The four criteria just discussed may be helpful in identifying clear cases of canonical agreement. Once identified, annotators may choose to ignore these formatives completely in the gloss. However, as the connection (or lack of) between overt agreement morphology and the presence/absence of free pronouns continues to be a focus of research, annotators may decide to gloss agreement morphology as an optional layer of detail in their annotations. For glossing canonical agreement, the same symbols for person/animacy are used, but without the <pro> symbol. The only addition to the inventory of forms is that the number <3> is required to indicate ‘agreement with a controller in the third person’ (with <pro> and <np>, third person is the default value, hence receives no additional gloss, cf. 2.2 above). In the interests of simplicity, and given the marginal interest of agreement within the GRAID-framework generally, we do not include a function gloss with the agreement marker. For example, the verbal agreement in the Turkish example could be either ignored completely, or glossed as follows:

(20) TURKISH

niçin gel-di-n?

why come-PST-2SG

0.2:S other v:pred-2 (‘verbal predicate + agreement with second person argument’)

‘why have you come?’

Free pronouns, i.e. prosodically independent items with word-order freedom, will always be glossed in a GRAID annotation, using the <pro> gloss for form. This is true even for languages (like English) where the free pronouns are obligatory in first and

second person clauses. While canonical agreement, and free pronouns, generally pose few difficulties for annotators, the correct analysis of bound or clitic pronouns is often highly contentious, both from a theoretical as well as a practical glossing perspective. We deal with some of these issues in the next section.

3.2.3 Bound pronouns vs. agreement

According to the first criterion mentioned above, that of morphological boundedness, clitic or bound pronouns are evidently closer to canonical agreement than free pronouns are. But they often differ from the latter in terms of the other three criteria, leading to conflicting results in their classification. There is a great deal of cross-linguistic variation in this respect, and annotators should be wary of across-the-board solutions. In what follows, we present some test cases, and propose guidelines for glossing. We will also take up the issue of so-called clitic-doubling, and suggest how it may be accounted for in GRAID annotations.

Let us begin with a rather clear example of a prosodically bound element that is nevertheless functionally equivalent to a free pronoun. The language is Central Kurdish (Indo-European, Iranian; North Iraq). All finite verbs in the present tenses carry canonical affixal agreement with S or A. But in addition, there is also a set of pronominal 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 (21a). If the latter is not present, the corresponding clitic pronoun attaches to the left-most constituent of the VP, in (21b) the negation prefix, giving the impression of an infix element:

- (21) SORANI
- 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 (22):

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

The direct object clitic can be considered to be pronominal, rather than agreement, because of its non-obligatoriness (it is not required in all clauses of this type). However, the distributional facts 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, annotators should be aware of the fact that argument cross-referencing systems are often construction specific, rather than language-specific, and annotators may well need to define their annotation processes for distinct constructions.

The bound pronouns of Sakapultek Maya show quite different properties (data from Du Bois 1987). Sakapultek has ergative alignment and strong head-marking tendencies. S, A and P arguments of the 1st and 2nd person are obligatorily realized as affixes on the verb:

- (23) 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.')

However, it is also possible for free pronouns to co-occur with the pronominal affixes. Du Bois suggests 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 pronominal expressions is extremely rare and pragmatically restricted to contrastive contexts (cf. Yup'ik discussed below). The situation with 3rd person arguments is similar:

- (24) 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.’

The paucity of free pronouns, and their restriction to certain pragmatically determined contexts, makes it reasonable to assume that the affixes in question, despite being morphologically bound and syntactically obligatory, should nevertheless be treated as pronouns, rather than canonical agreement.

This analysis immediately raises the issue of how the (rare) instances of free pronouns are to be glossed (often misleadingly referred to as “clitic doubling” in the literature): If the pronominal clitics are given argument status, then glossing the free pronouns as realizations of the same argument would run counter to very fundamental principles of most mainstream syntactic theory. Furthermore, such pronoun-doubling potentially violates binding conditions. But these are theory-internal stipulations which are not of primary concern here. Our position is to take such examples of double exponence at face value: as a discourse-motivated doubling of the reference to the indirect object. As such, it is of interest to GRAID annotators, because it effectively **increases the discourse saliency** of a particular argument within the clause. Thus it can be seen as taking the expression of the argument one step beyond the normal maximum, the ‘full NP’ option. We do not think that the **potential availability** of double exponence in the language would warrant a wholesale treatment of the clitics as canonical agreement, because the facts of discourse usage make it quite clear that this is actually quite rare. Those who find this solution unpalatable may opt for a different glossing alternative. For example, the free pronouns could be given the function gloss <dt> ‘dislocated topic’, or the function gloss <other>. Whichever option is chosen for Sakapultek, it would not have a very significant impact on the overall profile, as such clauses are rare in actual discourse.

Another example of bound pronouns displaying similar properties to free pronouns is found in the polysynthetic language Yup’ik, discussed in Mithun (2003). Mithun shows that bound pronouns in Yup’ik do not allow for an indefinite reading of the P argument; instead, an alternative construction is employed in order to express the same state-of-affairs (Yup’ik, Mithun 2003:251-252):

- (25) 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 general according to Mithun (2003), Yup'ik bound pronouns behave very much like English free pronouns in that they occur only in contexts where their referent is identifiable. In Yup'ik and other languages with bound pronouns on the verb, these are often the only means of expression of participants, differing in this respect from agreement markers:

- (26) YUP'IK (Mithun 2003:243)
- a. *arulaiqarluta*
 arula-ir-qar-lu-**ta**
 be.in.motion-NEG-briefly-SUBORD-1PL
 '... **we** stop briefly'
- b. *nayugaqurlaput*
 nayur-qaqur-la-**put**
 observe-intermittently-OPTATIVE-1PL/3PL
 'and watch **them** for a while?'

What makes Yup'ik different from English is the possibility to establish a new discourse participant with a lexical NP and a subsequent pronoun **within the same clause**:

- (27) YUP'IK (Mithun 2003:247)
- Yuut** piterrlainnayuitut*
 yuk-t pi-te-rrlainar-yu-ite-u-t
 person-PL thing-catch-constantly-HAB-NEG-INTR.INDIC-3PL
- '**People** don't always catch game.'

For the glossing of such contexts, the same set of options is available as was introduced above for Sakapultek. Note that GRAID glossing captures the distinction bound vs. free, and also captures person distinctions. This means that even after a particular glossing decision has been taken (e.g. to gloss the bound pronouns as <pro>), the individual

items are still recoverable and distinguishable from the free pronouns by virtue of the dash linking the bound pronouns to the <pred>-gloss.

Finally, consider a particularly well-known example, that of clitic pronouns in Spanish (data from Pineda and Meza, Undated), illustrated in the following examples:

(28) SPANISH

Juan muestro el catálogo a María
 J. show:PST:3S the catalogue to María

‘Juan showed the catalogue to María’

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

(29) SPANISH

- a. *Muestra=lo a María*
- b. *Lo=muestra a María*
- c. *Muestra=le el catálogo*
- d. *se=lo=muestra*
- e. *se=lo=muestra a María*

In the examples (a) and (b), only the THEME is pronominalized (the clitic *lo*) while the RECIPIENT remains as a full PP (*a María*). In (c), on the other hand, the RECIPIENT is pronominalized (*le*), while the THEME is a full NP. In (d), both the RECIPIENT (*se*) and the THEME (*lo*) have been pronominalized. The same is true of (e), but here the RECIPIENT also occurs, seemingly redundantly, as a full PP in the clause (*a María*). This example would appear to be a case of multirepresentation, hence bringing the clitic pronoun closer to the agreement pole of the cline. And indeed, some researchers do in fact consider the Spanish clitics to be agreement morphemes, which would take them outside the purview of GRAID annotations in the narrower sense.

However, we consider the real issue is not that of multirepresentation, but of obligatoriness. In none of the Spanish examples above is the pronominal clitic a grammatical necessity in the clause; its presence is determined by discourse considerations. Hence in (e), the indirect object clitic *se=* could have been omitted. Furthermore, the possibility of multirepresentation is semantically conditioned: it is only possible with animate referents and hence the clitics express - to some degree - descriptive content, a typical

property of pronouns rather than agreement markers. In other words, despite the possibility of multirepresentation of the object, the clitics in questions are to be analysed as bound pronouns rather than agreement. Our recommendation in such cases is therefore to gloss both the clitic pronoun and the NP-argument, yielding for example the following:

(30) SPANISH

se=lo=muestra *a María*
 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’

3.2.4 Summary of bound pronouns

As mentioned, the glossing of canonical agreement, and of free pronouns, poses few difficulties for annotators. The problems occur in the intermediate zone of bound pronouns / pronominal affixes / clitics. The four criteria above may be useful in identifying the phenomena and aiding annotators in their decisions. Annotators should also be aware that there are likely to be differences in the way the language treats cross-reference markers for the first and second person markers, and how it treats third person markers. Furthermore, languages regularly have distinct systems of argument cross-referencing for different construction types, so solutions may need to be construction specific, rather than language specific, and the relevant constructions need to be defined in the accompanying documentation.

Given the extent of the controversies surrounding the analysis of bound pronouns in the literature, it is inevitable that glossing them will raise certain problems. Determining the extent of variation in referential strategies, and the nature of the factors determining them, is the objective of GRAID annotations rather than a heuristic for decisions on glossing. But it is precisely by undertaking the glossing, and by comparing the results obtained through different glossing decisions that the GRAID analysis will contribute towards resolving borderline cases, and identifying commonalities.

3.3 Reflexive and reciprocal pronouns

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 personal pronoun (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. These strategies may also then be extended for use with 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 make 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>. We would extend this recommendation to reciprocal pronouns, which may also be glossed using the symbol <refl>.

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*), the decision on whether to code a reflexive pronoun as an argument must be left to the discretion of the annotator. 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.

3.4 Argument positions with non-finite predicates

With a number of non-finite predicates, for instance participial, infinitival or converb constructions, the overt expression of the highest-ranking argument (generally subject) is categorically blocked, and so this argument cannot be overtly realized in the clause. The question therefore arises as to whether in such cases a zero-glossing of the argument is appropriate. In Bickel (2003), the position is taken that where an argument position is systematically blocked, then no argument position is available and the corresponding argument cannot be glossed. This is a reasonable position, but it does of course carry

certain consequences. Consider for example the semantically very similar bracketed clauses in (31a) and (31b):

(31) ENGLISH

- a. *I promised my mother [to sell the motorbike]*
- b. *I promised my mother [that I would sell the motorbike]*

If we adopt the position that syntactically unavailable argument positions should not be glossed, then the two clauses would be glossed as having a different numbers of arguments (in (31a) just <np:P>, coupled with the <vother:pred> gloss, while in (31b) we would have two arguments: <pro.1:A> and <np:P>). If, on the other hand, we wish the gloss to more closely reflect underlying thematic relations, then we could gloss a zero-argument in the complement clause of (31a), i.e. <0.1:A> (essentially corresponding to PRO of Government and Binding Theory). In this manner, the argument glossing would reflect more closely the argument structure of (31b). It would, however, involve a rather radical departure from the surface forms of English, in that it suggests an argument in a position where no argument can ever occur. Thus our general preference is not to gloss an argument here, and to gloss the non-finite verb form *to sell* with <vother>, indicating that it does not allow the full range of arguments to be expressed.

Whichever solution is adopted, it will have certain consequences for the ratio of arguments to predicates (Bickel’s (2003) ‘Referential Density’), though the effect will of course depend on the overall frequency of such constructions in a particular language. If a language makes very widespread use of non-finite predicates, that could skew its profile when compared to other languages. This is, in itself, actually an interesting question,⁸ which GRAID annotations can contribute to answering. The only sound empirical procedure for such languages appears to be to test both annotating options (i.e. counting syntactically unavailable argument positions vs. not counting them) in order to determine how significant the effect would be.

3.5 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

⁸ One could in fact speculate whether languages like English, generally considered to have a high referential density, compensate for the high density of pronouns in discourse by making extensive use of non-finite constructions lacking overt subject pronouns. This merits closer investigation.

own internal predicate-argument structure. This raises certain problems for annotators, which we consider in this section.

The presence 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>. This gloss indicates that the following string is a complement clause. 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 (17) could be glossed as follows in GRAID:⁹

(32) 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
- 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):

(33) 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*:

(34) ?? *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 verb would lack a P-argument and would best be considered intransitive. Thus its subject

⁹ The example (32a) would probably be more naturally rendered with an expletive pronoun, as in *It surprised me that Shawn came to the party*. Glossing of such elements can be achieved in a number of ways, in the present example the simplest being <other:A>

would be glossed as S rather than A. Applying this option would yield the following gloss for (32b):

- (32b') *Irv believes that Harriet is a secret agent.*
 # np.h:S v:pred #cc:other other np.h:S cop np:pred

Center-embedded complement clauses pose similar problems to center-embedded relative clauses, discussed above in Section 3.1. Similar solutions can also be applied to complement clauses.

3.5.1 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):

- (35) ENGLISH
 a. *Irv believes [Harriet is a secret agent].*
 b. *Irv believes Harriet [to be a secret agent].*

In (35a), *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 (35b), on the other hand, there is good evidence for assuming 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:

- (36) 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 (35b) should be glossed as an object to *believes*. The possibilities for glossing the two clauses are given below:

- =(35a) # np.h:A v:pred #cc:P np.h:S cop np.h:pred
 or: # np.h:S v:pred #cc:other pro.h:S cop np.h:pred
 =(35b) # np.h:A v:pred np.h:P #cc:other vother:pred np.h:other

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*:

- (37) ENGLISH
- a. *It seems [that Harriet is a secret agent].*
 - b. *Harriet seems [to be a secret agent].*

In (37a) *Harriet* is an argument of the complement clause following the complementizer *that*. In (37b), *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 (37b) 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 §3.4 in connection with (31) 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. A case study of complex clauses in natural discourse from a language documentation project (Vera’a, Oceanic, Vanuatu) is provided under

<http://www.linguistik.uni-kiel.de/GRAID%20glossed%20text%20example.pdf>

References

- 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.
- 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.
- 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,.
- 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 pronouns: evidence from bininj gun-wok. *Sprachtypologie und Universalienforschung STUF* 52:255–281.
- Farrell, Patrick. 2005. *Grammatical relations*. Oxford: Oxford University Press.
- Forker, Diana. 2010. A grammar of Hinuq. Ph.d. dissertation, Universität at Leipzig, Philologische Fakultät at der Universität at 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 functional pronominal case: Evidence from Iranian. Presentation at the *Third International Conference on Iranian Linguistics*, Paris Sorbonne 3 (September 2009).
- 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, ??-?? London: School of Oriental and African Studies.
- Haig, Geoffrey, Stefan Schnell, and Claudia Wegener. 2011. Comparing corpora from endangered language projects: Explorations in typology with original texts. Ms., available on demand from the authors.
- Haspelmath, Martin. To appear. On S, A, P, T and R as comparative concepts for alignment typology. To appear in: *Linguistic Typology*.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195.
- Himmelman, Nikolaus P. 2006. Language documentation: What is it, and what is it good for. In *Essentials of language documentation*, ed. Jost Gippert, Nikolaus P. Himmelman, 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.
- 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.
- 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/>.
- Payne, Thomas E. 1992. *The twins stories: Participant coding in Yagua narrative*. 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, Rene, Balthasar Bickel, and Kristine Hildebrandt. 2010. The prosodic word is not universal but emergent. *Journal of Linguistics* 46:657–709.
- 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.
- 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 Özçalışkan, Psychology Press Festschrift Series, 543–555. London: Psychology 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.

4 Alphabetical list of GRAID symbols

#	clause boundary, inserted at left edge, one per clause
[...]	boundaries of embedded clause (optional, cf. p. 23)
0	‘zero’: argument position not filled by an overt referring expression
1 / 2	argument with 1st / 2nd person referent(s)
3	third person. This symbol is only used in the optional glossing of agreement morphology, discussed in Section 3.2.
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.
A (or: a)	transitive subject
adp	adposition
aux	auxiliary
cc	complement clause
cop	overt copular verb, in combination with some kind of non-verbal predicate complement, cf. Section 2.4.2
g	goal argument of a goal-oriented verb of motion, transitive or intransitive, may also extend to Recipient and Addressee (cf. p. 13)
h	NP has human referent(s), or refers to anthropomorphized referents
l	locative argument of verbs of location
dt	dislocated topic
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 (or: p)	transitive object
poss	possessor
pred	function gloss for the item that constitutes the predicate of a clause
pro	free pronoun in its full form (in contrast to <-pro>, or <=pro>)
rc	relative clause
refl	overt reflexive or reciprocal pronoun, cf. Section 3.3
S (or: s)	intransitive subject
v	lexical verb as the form element of a predicate
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>
