RefIND — Referent indexing in natural-language discourse

Annotation guidelines v1.1

Nils N. Schiborr [a] Stefan Schnell [b] Hanna Thiele [c]

15th March 2018

[a] University of Bamberg
nils-norman.schiborr@uni-bamberg.de

[b] University of Melbourne
stefan.schnell@unimelb.edu.au

[c] University of Bamberg
hanna.thiele@uni-bamberg.de

Table of Contents

1 Introduction **2**
2 Basic concepts **3**
2.1 Objects of RefIND annotations: what are discourse referents? 3
2.2 Basic format and content of annotations 3
3 Annotation guidelines **5**
3.1 Referential, co-referential and non-referential expressions 5
3.2 Specificity and individuation of referents 5
3.3 Ontological classes entities 6
3.3.1 Spatial and temporal referents: specific points in space and time, time intervals, areas 7
3.3.2 Non-physical entities: abstract concepts, states-of-affairs, speech acts, mental states 8
3.3.3 Specificity and levels of reality 9
3.3.4 Non-specific states of affairs 10
3.4 Newness of referents in discourse 11
3.5 Referentiality features in different morphosyntactic constructions 12
3.5.1 Incorporated nouns 12
3.5.2 Nominal predicates 12
3.5.3 Adnominal modifiers 13
3.5.4 Conflated objects and other arguments 13
3.5.5 Dummy subjects and dummy objects 14
3.5.6 Standards of comparison 14
4 Further information **14**
4.1 Use of RefLex 15
4.2 List of discourse referents 15
5 Research context and rationale **16**

References **19**

1 Introduction

RefIND stands for REFERENT INDEXING IN NATURAL-LANGUAGE DISCOURSE. It resembles a set of corpus annotation conventions to address research questions in the area of reference and discourse structure. As such, it is similar to the RefLex annotation scheme developed by Arndt Riester and Stefan Baumann (Riester & Baumann 2017), and we incorporate part of the RefLex scheme into our RefIND annotations. While RefLex glosses resemble explicit labels for different information statuses, like 'given' or 'new', RefIND uses mere indices, unique identifiers for each discourse referent. Consider the following example from a biographical account in the Kent variety of English:

(1) Multi-CAST `english_london01a_007` (Schiborr 2015)

And	*the*	*three*	*children,*	…	*mi*	*father*	*and*	*his*	*two*	*sisters,*
			001		002	003		003		004

were	*put*	*into*	*an*	*orphanage.*
				005

There are six referring expressions in this one sentence, and each of these receives a three-digit referent index. Expressions picking out the same referent receive the same index; thus both expressions for the narrator's father, the NP *mi father* and the possessive determiner *his*, receive the index ⟨003⟩. Reference to the narrator, their father, his sisters, the whole group of children and the orphanage all constitute distinct discourse referents, as will be explained in more detail in § 3 below, and therefore all receive distinct indices. Once texts are fully glossed, we can determine for all referents in a given stretch of discourse where they are mentioned for the first time and what the relationship is between multiple mentions of the same referent. Thus, in contrast to the RefLex scheme, referentiality-related information cannot simply read off RefIND annotations, but becomes accessible only via more complex analyses. On the other hand, RefIND is structurally simpler yet captures more complex information. We explain this in more detail in § 5 below.

In what follows, we first outline the basic concepts of RefIND in § 2, defining its object of annotation and the annotation architecture in conjunction with GRAID (Haig & Schnell 2014). In § 3, we turn to the relevant criteria for the major decision RefIND annotators have to make, namely whether a given referring expression construes a new discourse referent or not. In addition to referent indices, we do make use of a subset of RefLex glosses, and we list discourse referents in a separate list, containing further semantic and other information; this will be explained in § 4. In § 5 we provide an overview of the wider research context for which RefIND-annotated corpora are usable, as well as the basic logic of RefIND annotations and analyses.

2 Basic concepts

2.1 Objects of RefIND annotations: what are discourse referents?

RefIND targets the linguistic expressions of what we call DISCOURSE REFERENTS. In a narrative, for instance, discourse referents are essentially the characters and objects that a narrator will introduce and talk about (Du Bois 1980: 204). The major question is whether a referring expression is referential or not. We adopt a relatively simple definition proposed by Du Bois (1980):

> A noun phrase [*SST:* more generally, a referring expression] is *referential* when it is used to speak about an object as an object, with continuous identity over time. (Du Bois 1980: 208)

Hence, the major characteristics of discourse referents is that they can be talked about[1] and that they are TRACKABLE throughout discourse, that is the possibility to identify *the same* entity throughout a discourse.

It should be made clear here that what speakers 'identify' and 'track' in discourse are not real-world or fictive objects, but *representations in discourse*. Discourse referents in this sense are construed by speakers to refer back to and talk about. Crucially, this includes reference to unreal entities that speakers posit as 'existent' in the reality construed in a narrative or other type of discourse (cf. Gundel 1985: 101–102 for an example of a topical referent of which the speaker asserts that it does not exist). Our conception of discourse referents thus essentially draws on FAMILIARITY THEORY rather than theories of presuppositional semantics; see Baumann & Riester (2010: 1) for a short overview.

More importantly, our approach here is — like the RefLex approach — a 'data-oriented' one, aimed primarily at corpus annotation rather than semantic or philosophical theorizing. For a more fully-fledged linguistic theory of reference and an overview of the philosophical background and discussion, see Abbott (2010).

2.2 Basic format and content of annotations

RefIND annotations consist simply of multi-digit numerical glosses that uniquely identify a discourse referent and are associated with every expression of each respective referent. RefIND annotations are undertaken on richly annotated corpora that already have GRAID annotations (Haig & Schnell 2014). In terms of corpus development and annotation, we hence create a layered structure of different levels of annotation, where a bare transcription is

[1] This is presumably related to a pragmatic consideration on topichood, so that one can say that only discourse referents can ever be topics, although they have to meet further conditions to be acceptable as topics, see Gundel (1985: 90) for a discussion of this problem.

first supplemented by a free translation, then enriched through morphological glossing. To this complex of standard annotation levels GRAID annotations are added, and RefIND supply yet another layer of glossing (cf. Haig & Schnell 2014).

The referent indices are numbers, three or four digits long. They are entered in a 1-to-1 cardinal relationship with GRAID glosses for referring expressions, which in turn are aligned with word form glosses; note however that both GRAID and RefIND target entire phrases. With this method, we create consistently ordered layers of annotation for different levels of linguistic representation. The following example from the Vera'a corpus in the Multi-CAST collection (Haig & Schnell 2015) illustrates this:

(2) Multi-CAST `veraa_anv_001-003` (Schnell 2015)

	qōñ	*vō-wal*	*e*	*ruɯa*	*mē*	*=n*
	day	NUM-one	PERS.ART	HUM:PL	DAT	=ART
##	np:other	rn	ln	np.h:pred 001 r-new	rn	=rn

gunu	*-ruō*		*duru*	*=m*	*'ōgo*	*'ōgo*
spouse	-3DL		3DL	=TAM1	stay	stay
rn_np.h	-pro.h:poss	##	pro.h:s 001	=ln	v:pred	rv

vaa-van		*n=*	*reñe*	*anē*	*ne*	*wotoqtoqo*	
RED-go		ART=	woman	DEM1.A	AOR:3SG	pregnant	
other	##	ln=	np.h:s 002 r-bridging	rn	ln	v:pred	##

	ne	*visis*	*ēn*	*ni'i*	*reñe*
ZERO	AOR:3SG	deliver	=ART	small	woman
0.h:a 002	ln	v:pred	ln	ln	np.h:p 003 r-new

'One day, (there was) a couple. They two stayed [i.e. 'as time went by'], the woman got pregnant and gave birth to a little girl.'

At the beginning of this folkloristic narrative, a married couple is first introduced in a predicative NP and glossed ⟨001⟩. Its subsequent mention is again glossed ⟨001⟩ again. The following subject refers only to the woman, and thus receives a new index ⟨002⟩. The subject of the following clause has the same referent ⟨002⟩, and then the object of that clause introduces a new referent, the baby girl ⟨003⟩, and so forth. In this way, entire texts are annotated for all instances of reference to discourse entities, thereby allowing for detailed analysis of referent introduction and tracking.

As demonstrated in this example, first mentions of a referent receive an additional (simplified) RefLex tag, specifying additional referential properties of the referent beyond its discourse newness. Thus, the couple receives the RefLex tag ⟨`r-new`⟩, since it is a brand-new referent at the very beginning of the discourse. The mention of the woman creates a new discourse referent, but since the referent picks out an individual that is part of the couple, it is inferable from context, a so-called bridging anaphor (Baumann & Riester 2010: 3), therefore receiving the tag ⟨`r-bridging`⟩. The new-born baby is then again brand-new.[2]

The hierarchical ordering of annotation layers then enables complex analysis of combined annotation layers, as will be explained in § 5 below.

3 Annotation guidelines

3.1 Referential, co-referential and non-referential expressions

The central question to answer during RefIND annotations is whether a given referring expression constitutes a new referent in discourse or not. If it does constitute the first mention of a discourse-new referent, the expression receives a new index. If it has a referent already introduced at an earlier point, it receives the exact same index used previously for that referent. Where an expression is considered non-referential, it does not receive an index.

Distinguishing between referential and non-referential expressions is a major challenge, and the following guidelines provide criteria for answering this question. At the same time, these guidelines for future annotators reflect the decisions that we have made so far in annotating the current set of corpus data, and thereby also document our annotation practices.

3.2 Specificity and individuation of referents

The clearest instances of discourse referents involve reference to a *specific individuated physical real-world entity* located in a specific spatial and temporal dimension. Archetypical examples are real people and things referred to in biographical narratives that depict a specific period of time at a specific location. Example (3.4) is such a case. The 'reality status' of referents will be discussed in § 3.3.

In addition to these typical cases, we also consider those non-specific entities discourse referents where they are taken up again in subsequent discourse. These often occur in irrealis contexts, and an example will be discussed below. But we also include generic or class referents, since these

[2]One may of course argue that the introduction of *a* baby is inferable at this point in the narrative; however, we understand the specific reference to that *particular* girl to not be inferable, and thus gloss her as a (brand-)new referent.

can likewise be taken up and tracked through subsequent discourse. The following example is similar to Lambrecht's (1994) example *the whale*:

(3) ***The pygmy sperm whale*** *also has a small spermaceti organ . . .*
The Wikipedia Corpus, 'Spermaceti organ'

We exclude as non-referential those nominal expressions that merely evoke a class of entities to express a particular property or class membership without creating a trackable entity in discourse. Often, these expressions function as nominal predicates or predicate nouns, as in (4):

(4) *He is* ***a doctor****.*

Excluded are also instances of non-specific reference where a nominal expression designates a class of entities, but neither establishes the entire class as a referent in discourse nor evokes a particular referent as a member of the class that would be trackable through discourse. Examples typically involve conflated objects (e.g. *He was wearing* **glasses**) and near-idiomatic pre-fabricated expressions like (5):

(5) *We went to* ***the pub*** *on Friday night.*

Here, *the pub* neither refers to the class of pubs nor to a specific entity. The main point here is that in this particular context, the expression *the pub* does not trigger the interlocutor to identify a referent, since it is irrelevant. Rather, the entire expression *going to the pub* conventionally activates a particular event frame with its associated happenings and activities, and that is all that is relevant here.

3.3 Ontological classes entities

While references to people and things are the most common type noted so far in our corpora, speakers regularly make reference to all sorts of concrete but also more abstract entities, all of which can be trackable discourse referents in the sense that subsequent pick out the same entity. The following is a list of the less typical ontological types of discourse referent:

- spatial entities (e.g. locations),
- temporal entities,
- states-of-affairs,
- speech acts, and
- mental states (e.g. thoughts, ideas, etc.).

In the following, we collect some comments on these ontological types.

3.3.1 Spatial and temporal referents: specific points in space and time, time intervals, areas

We include spatial and temporal references, as long as these expressions create a trackable discourse referent. Thus, where a spatial expression identifies a clearly delimited location, this can qualify it as a discourse referent. This includes confined areas or places that can be taken up again in subsequent discourse. These are typically referred to by for instance place names or local noun phrases like *the beach* where it clearly identifies a particular area (as in many Oceanic cultures). Included here are also specific routes between two places, like *the path from A to B is long*.

Issues arise with the treatment of spatial relations, like *inside*, *next to*, *in front of*, and so on. In particular in languages where these are expressed nominally, hence as potentially referential expressions, annotators will have to decide whether they make reference to a trackable entity in a given discourse context, as is the case in the following example from the Multi-CAST Northern Kurdish corpus:

(6) Multi-CAST `nkurd_muserz01_059` (Haig & Thiele 2015)

Min	*sandiq-ek-î*	*hesin*	*ji*	*deniz-ê*	*deran-îye,*
1SG.OBL	box-INDEF-EZ	iron	from	sea-OBL	pull-3SG.PERF
hundur-ê	*wî*	*tijî*	*zêr*	*û,*	*xezîne*	*ye.*
inside-EZ	3SG.POSS	full	gold	and	treasure	COP.3SG

'I pulled a chest of iron out of the sea, it [lit. 'its inside'] is full of gold and treasure.'

As a rule of thumb, in view of our definition above, it is generally more likely that INTRINSIC spatial relations as in (6) constitute discourse referents as opposed to RELATIVE spatial relations, which are usually more ephemeral and merely construed *ad hoc* in a given discourse context, and therefore do not establish a stable relation between a linguistic expression and a construed spatial entity.

Similarly, points in time or temporal intervals, as well as phases of the day, days of the week, months, and so on, are generally treated as discourse referents as long as they are delimited and trackable. For example:

(7) Multi-CAST `veraa_hhak_132` (Schnell 2015)

Dirōl=m	*mi'ir*	…	*diñ*	*ēn*	*ma'ava*	*ne*	*ma'ava*
3TL=TAM1	sleep		reach	ART	morning	TAM2:3SG	morning

'They slept until (the next) morning, and when it was dawning…'

Here, *ma'ava* 'morning' refers to a specific point in time, and it would be possible to refer back to it in subsequent discourse. While reference to specific locatable temporal intervals or points in time are included in RefIND, we exclude mentions of non-specific and unbounded times, like *in the past / future*, and unspecified points in time like *one day* — where the exact 'location' of that

time is irrelevant — as well as repeated non-specific temporal relations, like *in the morning*, *on Wednesdays*. See the following example from the Multi-CAST Teop corpus:

(8) Multi-CAST `teop_iar_045` (Mosel & Schnell 2015)
Peho vuri me paa sue ...
one day and TAM3 say
'One day she said, ...'

In the discourse following (8), no reference is ever made back to that day where the woman spoke these words, and the day is also not relevant for the story in general. In essence, these phrases appear to stand in for specific frames of reference and could hence just as well be translated as 'what happened next was...'.

A possible test for the referential status of spatial and temporal entities is their ability to be substituted by spatial or temporal proforms, analogously to the possibility of using anaphoric pronouns with non-specific things or persons. In the example *We slept until midday, and then we went out for breakfast*, for instance, *then* would refer to the same point in time as *midday*. It is even conceivable that *(back) then* could refer back to a phrase like *at a time very long ago*, which would mostly be used instead of *once upon a time* or similar expressions, and not designate a referable point of time.

More often than not in actual discourse data from under-studied languages, this test of substitutability alone may not allow annotators to arrive at a decision. This leaves many cases of spatial and temporal relations in a gray area. Annotators may leave these cases undecided by using the dummy glosses ⟨`se`⟩ and ⟨`te`⟩ in place of a RefIND index to mark problematic instances.

3.3.2 Non-physical entities: abstract concepts, states-of-affairs, speech acts, mental states

Non-physical entities generally constitute possible discourse referents. However, we exclude abstract concepts such as *hatred*, *mood*, and *justice* where these are mentioned as mere concepts, not relating to a person's mental state. An individual's emotional state, for instance *love*, in most cases does not represent a discourse referent, as in examples like *She is in love*.[3] In some cases, however, such emotional states and associated behaviors may be treated as discourse referents, in that they may be taken up again in subsequent discourse:

(9) *... and he'd seen her love and revelled in it, ...*
BNC-JY4_[her love]

This appears to be an area of much uncertainty, and in most cases instances of the later type will only reveal themselves once they are subject of anaphoric

[3]This seems to be an idiomatic expression similar to the ones discussed in § 3.5.4 below.

reference.[4] We recommend conservative annotation practices and indexation of only those cases that are beyond doubt, which generally leads such mental and emotional concepts to be excluded.

Similar abstract states are *alive/life* and *dead/death*. The following example involves a possessive construction encoding the association of this specific state with a specific individual:

(10) Multi-CAST `veraa_hhak_120` (Schnell 2015)

Kamadu	*me*	*vus*	*wal*	*ēn*	*es*	*nō-m*	*sa*	*qiri*
1DL.EX	FUT	kill	once	ART	life	POSS.DOM-2SG	EMPH	today

anei!
DEM4

'We will extinguish your life on this very day!'

Related to abstract states of this kind are states-of-affairs (SOAs) like events, which are likewise possible as discourse referents, and are generally included in RefIND annotations. As a rule of thumb, it is only actualized and temporally bounded SOAs — typically located prior to the time of speaking or some other reference time (RT) — that are referred to as an instance by a nominal expression that are considered referents. These instances typically involve nominalizations. Other instances of temporally bound specific SOA referents considered discourse referents are those expressed by so-called 'event nouns' like *concert*, *tournament*, *race*, and so on.

3.3.3 Specificity and levels of reality

In natural discourse, speakers regularly express states-of-affairs that are unreal or have not yet happened, and are thus IRREALIS. Descriptions of irrealis states-of-affairs can involve already known referents as participants, or they can involve participants that only 'exist' in the irrealis frame. Participants in both types of context can be discourse referents, as can be seen from the following set of examples:

(11) a. *When I finish high school, I will marry a handsome guy. We will have at least two kids.*

b. *After finishing high school, I will marry my boyfriend Tim. We already have two kids.*

(12) *He will have a good job after finishing law school, but he will also look after our children. But if I don't love him anymore after a while, I will dump him and ...*

[4] What is relevant here is whether a given concept can be the topic of a proposition, in which case it constitutes a discourse referent. Where an expression occurs as subject or object, it is interpreted as (potentially) topical. In (9), *her love* bears a participatory role.

This example involves two variant scenarios, so that (12) may be the continuation of either (11a) or (11b). While both versions in (11) are located in the future and as such hypothetical, the first involves a non-specific, but individuated would-be partner of the speaker, while the second involves the specific, real boyfriend *Tim*. Differences in specificity of the referent do not appear to matter at all for the conception of *discourse referent*. The more important point here is that an entity has been *construed* as a referent in discourse; whether it actually exists or not is irrelevant. This is also reflected in part in the discussion on topicality: Gundel (1985), for instance, shows that even referents of which a speaker explicitly states that they do not exist in reality can be a sentence topic (see Gundel 1985: 101-102), and all that matters is that a 'hypothetical' referent is plausible and familiar through discourse context.

Similar instances are attested regularly in corpora of narrative texts like folktales, often involving foreshadowing by the narrator or prophecies expressed in direct speech by characters within the story. From the discussion above it is clear that a foreshadowed character does constitute a discourse referent. More important is the relationship between the foreshadowing and the point at which the narrative arrives at the foreshadowed events, where the events are actually *at* reference time (RT). As a rule of thumb, we suggest that the foreshadowed and later actualized instances of reference are treated as two different discourse referents, so that the later encountered entity is also treated as discourse-new, thus receiving a new index. Foreshadowing may also involve reference to specific entities that are already established at RT where the foreshadowing takes place, in which case the same referent index would have to be used for both instances of reference.

The same complexities apply to cases involving complement clauses of desiderative and other verbs evoking a point in time posterior to this event. For instance, if a speaker says,

(13) *I was looking for* ***a horse****, . . .*

and then, at a later point in the narrative, continues with

(14) *. . . and when I finally found* ***a horse****, it was so very handsome, and . . .*

then we consider the two instances of *a horse* in (13) and (14) to have different referents precisely in the way described above.

Excluded from RefIND annotations are negative indefinites like 'nobody', as they are non-referential.

3.3.4 Non-specific states of affairs

The same considerations discussed in the preceding section may crop up in connection with reference to states-of-affairs, for instance when an individual is planning to do something. These appear to be more problematic than references to physical anticipated entities, since they are not bounded in time

and thus not well delimited and trackable. They are therefore generally not considered discourse referents, except for clear cases of anaphoric reference to these SOAs in the subsequent discourse (see §3.3.1 above for the same criterion regarding spatial and temporal reference). Consider this example:

(15) Multi-CAST `veraa_hhak_070` (Schnell 2015)

Ne	*revrev*	…	*van*	*lē=n*	*go-go'.*
TAM2:3SG	evening		go	LOC=ART	RED-hook

'When it was getting dark [they set off], (they) went fishing.'

3.4 Newness of referents in discourse

As outlined above, we consider a referent 'new' when it is mentioned for the first time in discourse, so that it is not given in the current discourse context. Crucially, the same applies to referents that are, in Prince's (1992) terms, *discourse-new* but clearly *hearer-old*, since they are part of the interlocutors' encyclopaedic knowledge, like the first mention of *the sun*. Similar considerations apply to referents inferable via 'frame semantics' (Fillmore 1982): consider for instance Hawkins' (1978: Ch. 3) example of the parts of a car being identifiable from the mentioning of the semantic concept 'car' (cf. also Lambrecht 1994: 91). The latter is considered a new discourse referent.

Similar to the inferability of referents via frame semantics are cases where a referent is associated with another referent through individual set membership. In these cases, one referring expression picks out more than one individual as a discourse referent. A common example is reference made to a group of people, like the three siblings in the English example above. The other referring expression picks out only a subset of this referent. Two scenarios are possible:

- **partial co-reference**: a set of individuals is initially referred to, either once or multiple consecutive times, and thereby a single 'discourse referent' is created. Subsequently, a subset of these individuals is picked out as the referent of some other referring expression, thus creating a new discourse referent. A typical instance of such a subsequent expression in English is *one of them;*
- **split antecedence**: two or more distinct discourse referents are mentioned initially, either once or multiple consecutive times. Subsequently, the individuals previously realized as multiple discourse referents are conjointly referred to by a single referring expression, thereby creating a new discourse referent. A typical instance of such a subsequent expression in English is *the two of them*, or, in many languages, a non-singular pronoun.

In both cases, the subsequent use of a referring expression creates a new discourse referent, that is, it establishes a new stable relationship between

linguistic expressions and a construed entity, the latter of which overlaps the one referred to initially, but is not exactly the same as it. The respective referring expressiosn receive new referent indices.

Once more, we stress that discourse referents are 'construals' of perceived or imagined realities that result from singling out elements from an unordered mass of elements. This is achieved via use of refering expression to point to these referents over a stretch of discourse.

In this sense, a discourse referent resembles a package, bundled up so it can be talked about and tied together by means of linguistic expressions. Whether the content of such a package consists of a single or multiple individuals is irrelevant from the point of view of reference. It is relevant only in so far as the content of the package may trigger the inferability of referents that are part of other 'reference packages'; compare the discussion of referents in Prince (1981), and the notion of bridging anaphors in Huang (2000).

3.5 Referentiality features in different morphosyntactic constructions

In the preceding sections, we have outlined our criteria for 'referenthood'. We now turn to instances of particular morphosyntactic constructions that typically involve non-referential nominal elements, such as incorporated nouns. In accordance with our definition above, we interpret *non-referential* as 'not evoking a trackable referent'. The types of instances described in the following hence essentially involve the same criteria for referenthood, and are merely intended to illustrate cases of non-referentiality and help annotators recognize them.

3.5.1 Incorporated nouns

We generally consider incorporated nouns to be non-referential, unless specific properties of these suggest otherwise, as discussed for instance for the Australian language Bininj Gun-wok (Evans 2002: 21–22).

There may be instances of incorporated nouns where a referent of the same concept is introduced in the subsequent context. For instance, if we encounter an example like *They went fish-hooking*, and the *fish* that have been caught are specifically mentioned afterwards, we treat this mention as the first mention of a new referent that is classified as 'bridging'.

3.5.2 Nominal predicates

Nominal predicates in non-verbal clauses expressing mere concepts are treated as non-referential and are thus excluded from glossing:

(16) Multi-CAST `veraa_multi-cast_isam086` (Schnell 2015)

di *=n* *ʼañsara*
3SG =ART person
'He is (a) human (being).'

Such nominal predicates attribute a property to the referent of the subject rather than evoking an entity. Compare also (4) from English above.

While existential construction with affirmative polarity have referential noun phrases in predicative function, as demonstrated in (2) above, referring expressions in the scope of negated existentials of the type *there was no X* or *an X did not exist* are not referential, and thus not counted as discourse referents and not glossed. See the following example from Teop:

(17) Multi-CAST `teop_iar_113` (Mosel & Schnell 2015)
Ahiki *ta* *tapeako,* *ae* *ta* *kaukau.*
not.exist NSPEC1.SG manioc AND1 NSPEC1.SG sweet.potato
'There was no manioc, and no sweet potatoes.'

3.5.3 Adnominal modifiers

Non-specific possessors that merely attribute a property to an entity are not treated as referential expressions. Typical examples from English include *a ray of light* or *a badge of honour*.

The same applies to otherwise embedded nominal expressions that function as attributes within higher-order referring expressions, like *a man* ***in a suit*** or *mushrooms* ***braised in red wine***, where neither *suit* nor *wine* are referential. The same applies to non-head elements of compounds, be these word-like or phrase-like. Typical examples from English would be ***chicken*** *drumsticks*, ***steel*** *bar*, and ***fish*** *bone*.

We also exclude instances of certain types of location nouns (common in Oceanic and Germanic languages) where these do not create a spatial referent, but merely specify the exact spatial relations between two entities (cf. §3.3.1 above).

3.5.4 Conflated objects and other arguments

Conflated arguments are likewise deemed non-referential. This includes conflated objects, such as in *wear glasses*, *don a suit and tie*, and *play the guitar*, and similar oblique arguments as in *go to the beach* and *go to the pub*, where in principle any *beach* or *pub* would do.

These nominals may often be the complement of so-called vector (or light) verbs or part of fixed idiomatic expressions such as *take a shower*, *have a shave*, and *kick the bucket*. In our view, object noun phrases in idiomatic expressions designate a concept that is not part of the meaning of the expression as a whole, and hence do not represent discourse referents. See Singer (2011) for a discussion of idiomatic expressions of this kind.

3.5.5 Dummy subjects and dummy objects

In some languages, certain syntactic argument positions need to be filled for purely structural reasons. Typical examples are expletive subjects in Germanic languages like English and German. In some languages of East Asia, we also find 'dummy objects' like *rice* and *books* that have to co-occur with verbs meaning 'eat', 'see', and 'read', but do not refer to specific participants being eaten or read.

Note that dummy subjects and objects differ fundamentally in their form and content: expletive subjects are essentially pronouns that neither bear conceptual meaning nor have a referent; dummy objects are lexical nouns that have conceptual meaning, but do not refer to a specific referent representing the concept.

3.5.6 Standards of comparison

NPs that designate a standard of comparison in expressions meaning 'like X' are often, but not always, non-referential, as in the following example from Teop:

(18) Multi-CAST `teop_iar_054` (Mosel & Schnell 2015)

Na	*potee*	*nana*	*bono*	*rupi*	*toa.*
TAM2	like	3SG.IPFV	OBJ.ART3.SG	egg	chicken

'(It) (looked) like a chicken egg.'

Iin this example, no *chicken egg* is introduced as a new discourse referent; the concept of a 'chicken egg' is merely evoked for the sake of simile with the object under discussion (a lump of earth).

Compare this to (19), where the standard of comparison is in fact referential:

(19) *You talk just like my brother.*

4 Further information

RefIND annotations are amended by two further sets of information: the RefLex tags for first mentions, and a tabularic list of referents. In Multi-CAST, the latter are part of the supplementary material accompanying the recording, annotation file, and metadata for each session.

For more information on our multi-language corpus collection Multi-CAST, please refer to our webpage at the Language Archive Cologne[5] and the *Multi-CAST structural overview* (Schiborr 2016), available there.

[5] `https://lac.uni-koeln.de/de/multicast/`

4.1 Use of RefLex

We combine RefIND annotations with a simplified version of the RefLex annotation scheme devised by Stefan Baumann and Arndt Riester (Riester & Baumann 2014). RefLex glosses are placed on additional annotation tier dependent on the RefIND tier, as demonstrated in § 2.2. We restrict the implementation the RefLex scheme to the first mentions of a discourse referent. As such, only the following RefLex glosses are relevant for RefIND (Riester & Baumann 2014: 3–4): *r-bridging, r-cataphor, r-bridging-contained, r-unused-unknown, r-unused-known,* and *r-new*, plus the optional feature +*generic*. We simplify this set of tags to the following three:

- ⟨`bridging`⟩: the referent is inferable from frame semantics, a previously mentioned scenario or situation, or is anchored to an already given referent that is expressed as an adnominal modifier (i.e. 'bridging-contained');
- ⟨`unused`⟩: a globally known entity (via encyclopaedic or cultural knowledge), e.g. *the sun*; and
- ⟨`new`⟩: a new referent not otherwise inferable or globally known.

We thus do not distinguish between known and unknown among the unused referents, a distinction that is often difficult to recognize anyway. Furthermore, the *bridging-contained* category is subsumed here under ⟨`bridging`⟩, and cataphors are not glossed as such at all. The relevant features of the latter types are still inferable from our layered annotation structure.

4.2 List of discourse referents

Every discourse referent recognized during RefIND annotations is entered into a list, comprised of the following information: (i) the three to four-digit referent index, (ii) a label or name for the referent, (iii) a short description of the referent, (iv) its semantic class and (v) relation to other discourse referents, and (vi) comments.

With regard to distinctions of semantic class, we are mainly interested in animacy ('ontological') categories. We distinguish the following:

- **human:** human beings and anthropomorphized non-human beings in fiction (e.g. animals in fables);
- **animate:** animals (not anthropomorphized);
- **inanimate:** things, inanimate physical objects;
- **body part:** body parts of human beings;
- **mass:** non-individuable masses like water, sand, etc.;
- **location:** physical locations, places, and areas;
- **time:** points or periods of time;
- **abstract:** emotions, thoughts, ideas, speech, etc.

We recognize the following relations to other referents:

- **set member of**: the individual referred to is a member of a previously introduced referent containing multiple members;
- **includes**: the referent embraces mutliple individuals that were previously established as separate referents;
- **part-whole of**: an inanimate object or mass is part of another object or mass previously referred to.

In the comments column, annotators may note any difficulties with identifying the referent in question or any other features that they think should be kept in mind, such as cultural information that users not familiar with the background of the language may require to understand the context of reference.

5 Research context and rationale

The research questions for which the GRAID+RefIND set of annotation procedures have been designed are situated in the area of research pioneered in the 1970's and 1980's by Wallace Chafe, Talmy Givón, and later John Du Bois. In short, our annotation scheme is intended to be used in research on discourse structure, its interaction with syntactic structure, and the choice of referring expressions.

This essentially comprises two main aspects of referentiality, namely how referents not previously mentioned are *introduced* into discourse, and how they are *tracked* through the subsequent discourse. The first question relates to postulated information-packaging aspects of discourse, like 'information pressure' and 'information density', and the use of special types of constructions — information that can gleaned from our annotations. The latter comprises well-known concepts of 'referental choice', involving dimensions of look-back, anaphoric distance, anaphora resolution, and so on.

The following are but a few examples of related, more specific research questions that GRAID+RefIND allows us to address:

- Which syntactic constructions (if any) are used specifically to introduce referents into discourse?
- Which syntactic functions and relations are involved in the introduction and tracking of discourse referents?
- Are different constructions and functions preferably used with human or non-human referents?
- What is the connection between antecedent distance and/or function, and choice of referential device?

RefIND annotations are intended to capture information that address these types of research question. In this regard, RefIND is very similar to RefLex, as

mentioned above. One difference between the two approaches to annotation is that in RefLex, information status features are directly glossed not only for discourse-new referents, but also for discourse-given referents. In RefIND, the information that a mention has a given referent will have to be derived from the fact that the index has an identical antecedent index somewhere in the preceding annotation.

Our two main reasons for favouring referent indices over the full RefLex scheme are as follows: in our experience, annotations with indices only is considerably quicker due the relative analytical simplicity of the considerations involved. For instance, for discourse-given referents, RefLex requires the annotator to make a distinction between *r-given* and *r-given-displaced*, where the latter is used for given mentions whose antecedent 'occurs earlier than the previous five intonation phrases' (Riester & Baumann 2014: 6). In RefIND, the same information is calculated during corpus analysis rather than during corpus annotation.

Moreover, where RefLex captures only the preceding five intonation units, we can calculate exact figures for antecedent distance. Although we concur with (Riester & Baumann 2014: 6) that five intonation units may be a reasonable distance at which changes in postulated activation are likely to take effect, as for instance in the form of referring expressions, the possibility to calculate exact figures offered by RefIND also enables re-assessment of relevant distances. Our calculations can be performed with different reference points, be they intonation units, clause units, total intervening referring expressions, mentions of other referents (of different semantic types, as noted in the list of referents), and so on.

The latter point is highlights another key difference, namely that RefIND referent indices *identify* discourse referents. The identification of individual referents allows us to determine whether one (type of) referent behaves differently from another, for instance with regard to antecedent distance. More basically, it allows us to describe global properties of a text in terms of the referential information contained therein, as makes it possible to calculate the absolute number of discourse referents (of different types) in the text, the referent density (Du Bois' 'information pressure'), or the ratio of type and token frequencies for discourse referents. These referent-oriented properties of whole texts have been claimed to be responsible for certain distributional patterns of referring expressions of different types in different works on referentiality and discourse structure, and RefIND annotated corpora thus allow for systematic assessment of these claims.

The following summarizes the properties and possibilities of RefIND annotations:

- identification of any mention of a referent in a given discourse, including its first mention;

- identification of the number of discourse referents and the number of their mentions in a given discourse;
- determining relationship between subsequent occurrence of the same index with regards to discourse structure, and to other indices.

Of particular interest for discourse structure are the first mentions of discourse referents, as they allows us to correlate all referent introductions with actual forms of reference, or even lexical items and phrases, as well as syntactic functions (and thereby to some degree syntactic constructions). Related questions concern preferences for referent introductions in terms of syntactic argument structure (Du Bois 1987, 2003a, b; 'preferred argument structure'), specialized referent-introducing constructions, for instance existential or presentational constructions (Lambrecht 1994; Abbott 1993), and questions about constraints on first-mention forms (pronoun-versus-noun phrase; see Bickel 2003 on Belhare). Moreover, the quantification of first mentions also provides a measure of the total number of discourse referents in any given discourse, which then allows us determine what Du Bois (1987) labels 'information pressure', later called 'referent pressure' in Du Bois (2003a, b), which is in turn relevant for different aspects of referent tracking and discourse structure.

The token frequency of each referent is connected to the thematic prominence (Lichtenberk 1996; Himmelmann 1997) of discourse referents, and may (arguably) influence the way in which discourse referents are initially introduced and subsequently established and mentioned in discourse; Bischoffberger & Schnell (2014) for critical assessment of these claims.

Lastly, we can determine the relationship between different instances of different mentions of a referent, and their respective formal and functional properties as registered in GRAID annotations. For example, we can calculate for every discourse referent the distance between mentions and related forms of expressions or syntactic function. This is connected to questions of referent accessibility as conditioned by the discourse context (Ariel 1988, 1990), or 'activation' (Chafe 1976, 1994), and 'lookback' (Givón 1983).

Considering that our initial experience suggests referent indexing to be comparatively straightforward, practically speaking, it appears to be a notably economic way to determine systematic information relevant for these long-standing research in corpora hitherto not considered. As such, we hope for a combination of GRAID+RefIND glossing to yield unprecedented findings in discourse structure and information management.

References

Abbott, Barbara. 1993. A pragmatic account of the definiteness effect in existential sentences. *Journal of Pragmatics* 19(1). 39–55.

Abbott, Barbara. 2010. *Reference*. Oxford: Oxford University Press.

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(1). 65–87.

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

Baumann, Stefan & Riester, Arndt. 2010. Annotating information status in spontaneous speech. *Proceedings of the Fifth International Conference on Speech Prosody* 100092.

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

Bischoffberger, Julia & Schnell, Stefan. 2014. *Thematic prominence and referential choice*.

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

Chafe, Wallace. 1994. *Discourse, consciousness, and time*. Chicago: The University of Chicago Press.

Du Bois, John. 1980. Beyond definiteness. In Chafe, Wallace (ed.), *The Pear Stories*, 203–274. Norwood, NJ: Ablex.

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

Du Bois, John. 2003a. Argument structure. In Du Bois, John & Kumpf, Lorraine & Ashby, William J. (eds.), *Preferred argument structure*, 11–60. Amsterdam: John Benjamins.

Du Bois, John. 2003b. Discourse and grammar. In Tomasello, Michael (ed.), *The new psychology of language*, 47–88. Mahwah, NJ: Erlbaum.

Evans, Nicholas. 2002. The true status of grammatical object affixes. In Evans, Nicholas & Sasse, Hans-Jurgen (eds.), *Problems of Polysynthesis*, 15–50. Berlin: Akademie Verlag.

Fillmore, Charles J. 1982. Frame semantics. In The Linguistic Society of Korea (ed.), *Linguistics in the morning calm*, 111–137. Seoul: Hanshin.

Givón, Talmy. 1983. Topic continuity in spoken English. In Givón, Talmy (ed.), *Topic continuity in discourse*, 343–364. Amsterdam: John Benjamins.

Gundel, Jeanette K. 1985. Shared knowledge and topicality. *Journal of Pragmatics* 9(1). 83–107.

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse)*. (`https://lac.uni-koeln.de/en/`

multicast/) (accessed 2015-12-30.)

Haig, Geoffrey & Schnell, Stefan (eds.). 2015. *Multi-CAST*. (https://lac.uni-koeln.de/multicast/) (accessed 2016-02-08.)

Haig, Geoffrey & Thiele, Hanna. 2015. Multi-CAST Northern Kurdish. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/multicast-northern-kurdish/) (accessed 2016-02-22.)

Hawkins, John A. 1978. *Definiteness and indefiniteness*. London: Croom Helm.

Himmelmann, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase* [Deixis, article, noun phrase: On the emergence of syntactic structure]. Tübingen: Niemeyer.

Huang, Yan. 2000. *Anaphora*. Oxford: Oxford University Press.

Lambrecht, Knud. 1994. *Information structure and sentence form*. Cambridge: Cambridge University Press.

Lichtenberk, František. 1996. Patterns of anaphora in To'aba'ita narrative discourse. In Fox, Barbara (ed.), *Studies in anaphora*, 379–411. Amsterdam: John Benjamins.

Mosel, Ulrike & Schnell, Stefan. 2015. Multi-CAST Teop. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/multicast-teop/) (accessed 2016-02-22.)

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Cole, Peter (ed.), *Radical pragmatics*, 223–255. New York: Academic Press.

Prince, Ellen F. 1992. The ZPG letter. In Mann, William C. & Thompson, Sandra A. (eds.), *Discourse description*, 295–325. Amsterdam: John Benjamins.

Riester, Arndt & Baumann, Stefan. 2014. *RefLex scheme*. Stuttgart / Cologne: University of Stuttgart / University of Cologne (doi:10.18419/opus-9011). (accessed 2018-03-03.).

Riester, Arndt & Baumann, Stefan. 2017. *The RefLex scheme — Annotation guidelines* (SinSpeC: Working papers of the SFB 732 14). Stuttgart: University of Stuttgart. (http://elib.uni-stuttgart.de/handle/11682/9028) (accessed 2018-03-01.)

Schiborr, Nils N. 2015. Multi-CAST English. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/multicast-english/) (accessed 2016-02-28.)

Schiborr, Nils N. 2016. Multi-CAST structural overview. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/multicast/)

Schnell, Stefan. 2015. Multi-CAST Vera'a. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST*. (https://lac.uni-koeln.de/multicast-veraa/) (accessed 2016-02-22.)

Singer, Ruth. 2011. Typologising idiomaticity. *Linguistic Typology* 15(3). 625–659.