Detailed annotation scheme for word order project

# Identification of ordered elements

The first element in the annotation scheme is the identification of ordered elements, i.e. arguments which are explicitly linguistically signalled. This step is **not** based on syntactic tests such as optionality, but on semantics: an element dependent on the verb is considered an ‘argument’ as long as the role that it plays in the clause is necessarily involved in the event denoted by the verb. This definition naturally includes, but is not exclusive to, elements traditionally considered arguments like subjects and objects. Consider the following:

[Tibetan Section of a University Library]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ཨ་ནི་ | འདི་གས་ | ཁྱེད་རང་ | གིས་ | གཡར་ | ཐུབ་ ཀྱི་ རེད་ | པས ། |
| a.ni | ‘di-gas | khyed-rang | =gis | g-yar | thub-kyi-red | pas |
| then | this-PL | 2HON-RANG | =ERG | lend | can-PRES-FACT | Q |
| then | these | you | | can lend? | | |
| ‘Then can you lend me these (books)?’ | | | | | | |

Here, *these* is the thing being lent, and *you* is the person doing the lending. In traditional terms, these correspond to direct object and subject. An act of lending naturally includes a lender, a thing to be lent, and a borrower. The borrower was not mentioned and therefore not considered in this study, but the other two elements are present and hence counted.

In the following example, there are still two elements considered, even though neither is a subject:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| བྱས་ཙང་ | མོ་རང་ | ལ་ | ཡ་ | ལག་རྟགས་ | གཅིག་ | བསྐུར་ | དགོས་ཡོད། |
| byas.tsang | mo.rang | la | ya | lag.rtags | gcig | bskur | dgos-yod |
| so | she-EMPH | =OBL | uh | gift | one | send | want-EGO |
| so | to her | |  | a gift | | want to send | |
| So I want to send her a gift. [The Untrustworthy Neighbour] | | | | | | | |

Here, *to her* is the recipient (traditionally the indirect object) and *one gift* is the thing being transferred (traditionally direct object). The giver (traditionally the subject) is not present, and instead the two objects are considered the two relevant elements.

In the following example, one of the elements being considered is not traditionally a subject or object at all, though may be called an oblique object:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ང | ས་ | ཁོང་ ཚོ | འི་ | རྩ | ར་ | བཞག་ མི་ ཡོང་ ། |
| Nga | s | khong-tsho | ‘i | rtsa | =r | bzhag-mi-yong |
| 1 | ERG | 3HON-PL | =GEN | side | =LOC | leave-NEG-PROSP |
| I | | at their side | | | | will not leave |
| I’m not going to leave it at their place. [The Untrustworthy Neighbour] | | | | | | |

Here, *I* is the putter (traditionally subject), and *at their side* is the location at which the thing is put (traditionally not a subject or object). But the act of putting necessarily implies that there is a location where something is put, and hence this element is still considered as one of the ordered elements being investigated.

For edge cases, I erred on the side of inclusion. This has resulted in the inclusion of some elements that may not be traditionally considered arguments, such as the following:

* The destinations **as well as sources** of verbs of movement are always included.

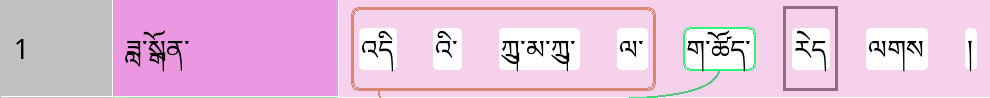
[The Lucky dream]



|  |  |  |
| --- | --- | --- |
| 41 | Dadrön: | From the market we arrived at the office. |

* In cost predicates, the entity whose price is being discussed is considered an element.

[Buying Vegetables in the Tromzikhang]



|  |  |  |
| --- | --- | --- |
| 1 | Dadrön: | How much is this tomato? |

* Directional demonstratives like ཕར་ *phar*, ཚུར་ *tshur*, ཡར་ *yar*, མར་ *mar* are not considered when they co-occur with a co-referring nominal in the same clause. However, they *are* considered when they do not co-occur with a co-referring nominal, as they do directly refer to a referent, e.g. *phar* for somewhere away from the deictic centre, *yar* for somewhere above the deictic centre. For example, ཕར་ *phar* in line 59 of the following is considered an argument:

[A Father Counsels His Daughter]

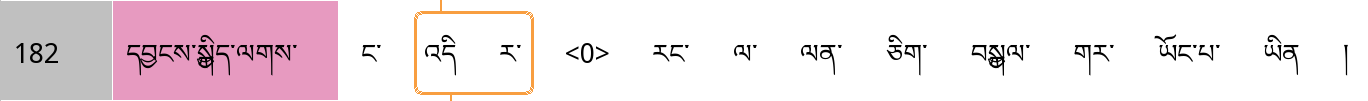
A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 58 | Father: | When she has difficulties, |
| 59 |  | Help that way (=her). |

* Positional demonstratives like འདིར་*‘dir*, ཕ་གིར་ *pha gir*, མ་གིར་ *ma gir*, འདི་པར་*‘di par* etc. are considered elements when they satisfy the above criteria as well. Thus, as sources or destinations of direction verbs, they are counted, such as the destination of *come* in line 182 of the following example; when just giving information on where an event took place, they are not counted.

[A Homemade Altar]

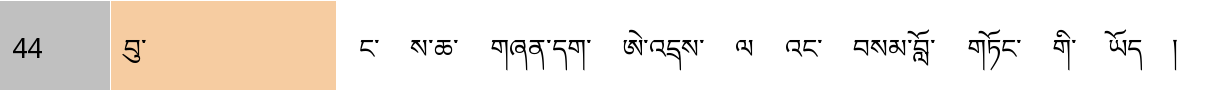


|  |  |  |
| --- | --- | --- |
| 182 | Yangki | I came here to deliver you a message. |

Similative demonstratives are not considered ordered elements unless in predicate position as they ordinarily carry manner information that is not specific to the verb.

Generally speaking, light verb nominals were discarded, as tend to be especially well integrated with the predicate (e.g. most of them cannot be modified, and they are often accompanied by other nominals.). However, they *are* counted if there is explicit grammatical implication that they are referential, such as the presence of modifiers, or noun phrases in the surrounding context that are clearly coreferential with it.

For example, the following line, there is no indication that བསམ་བློ་ *bsam.blo*  is treated as referential, and hence is not annotated as an argument whose ordered is to be studied:



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ང་ | ས་ཆ་ | གཞན་དག་ | ཨེ་འདྲས་ | ལ་ | འང་ | བསམ་བློ་ | གཏོང་གི་ཡོད། |
| nga | sa.cha | gzhan.dag | e.'dras | =la | =ang | bsam.blo | gtong-gi-yod |
| 1sg | place | other | SD | =LOC | =also | thought | LV-PRES.EGO |
| ‘I’m also thinking about other places.’ | | | | | | | |

One exception is the following line. Here, དགོངས་པ་ ‘thought’ is modified by a relative clause, an indication that it is more referential, and therefore it is counted within the scope of the study:



In some cases, an argument is expressed twice in the same clause. I will count those separately, leading to a ‘double count’ of a certain argument role. Consider the following example:

[The New Script]

A screenshot of a computer

Description automatically generated

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ང | ས་ | | // | ཁྱེད་རྣམ་ཚོ | | | | འི་ | གི་ |  |  |
| nga | s | |  | khyed-rnam.tsho | | | | 'i | gi |  |  |
| 1 | ERG | |  | 2-PL | | | | =GEN | =GEN |  |  |
| དགོངས་འཆར་ | | འདི་གས་ | | | | ང | ས་ | // | ཁོང་ | ལ་ | ཡ་ |
| dgongs.'char | | 'di=gas | | | | nga | s |  | khong | =la | =ya |
| suggestion | | DEM-PL | | | | 1 | ERG |  | 3HON | OBL | DM |
| བརྒྱུད་བསྒྲགས་ | | ཤིག་ | | | ཞུས་ | |  |  |  |  |  |
| relay | | INDEF | | | LV | |  |  |  |  |  |
| ‘I will, your suggestions I will relay to them.' | | | | | | | | |  |  |  |

Here, ངས་ *ngas* ‘1=ERG’ is uttered twice: Once before the theme argument ཁྱེད་རྣམ་ཚོའི་དགོངས་འཆར་འདི་གས་ *khyed.rang tsho’I dgongs ‘char ‘di.gas* ‘those suggestions of yours, and once after. A separate treatment is likely useful for an examination of the construction of syntax online, but this is set aside for this paper.

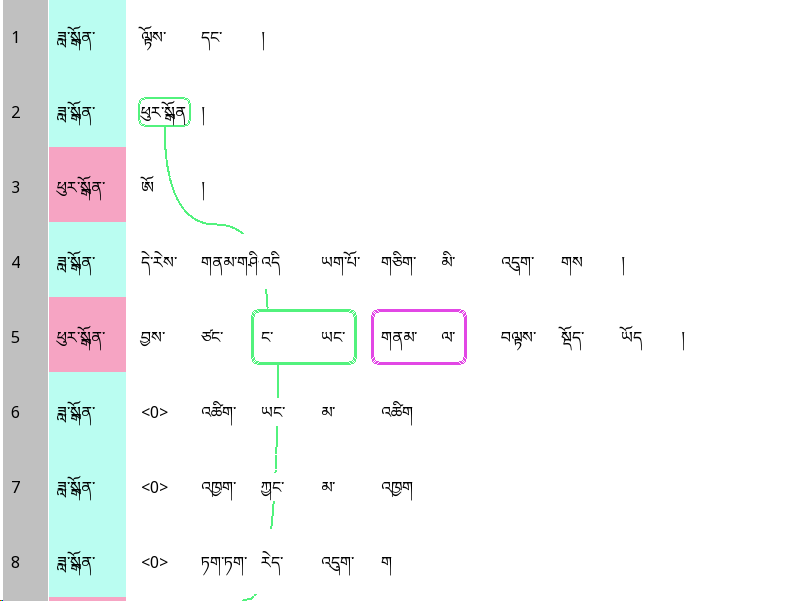
# Relationship of referent to context

## Previous and next mentions (noPrevMentions, noNextMentions)

Previous and next mentions are annotated using coreference chains, that is, chains of referring expressions that all refer to the same thing. This is done within Rezonator using the Track feature. In the following image, all coloured chains refer to the same entity. I code for the number of previous mentions within a previous ten-IU window (i.e. the 10 previous lines) and the number of following mentions within a following ten-IU window.

Refer to the example below, which shows a screenshot with the Rezonator interface, along with a translation in English. The green chain below groups together all expressions referring to the same person, Pudrön. The pink expression refers to the sky:

[A Lucky Dream]



|  |  |  |
| --- | --- | --- |
| 1 | Dadrön: | Look! // Pudrön. |
| 3 | Pudrön: | Oh. |
| 4 | Dadrön: | The weather is good today isn't it? |
| 5 | **Pudrön:** | **Then I also at the sky am looking.**  **Then I'm looking at the sky too.** |
| 6 | Dadrön: | It's not too hot |
| 7 |  | Not too cold |
| 8 |  | Just right, right? |

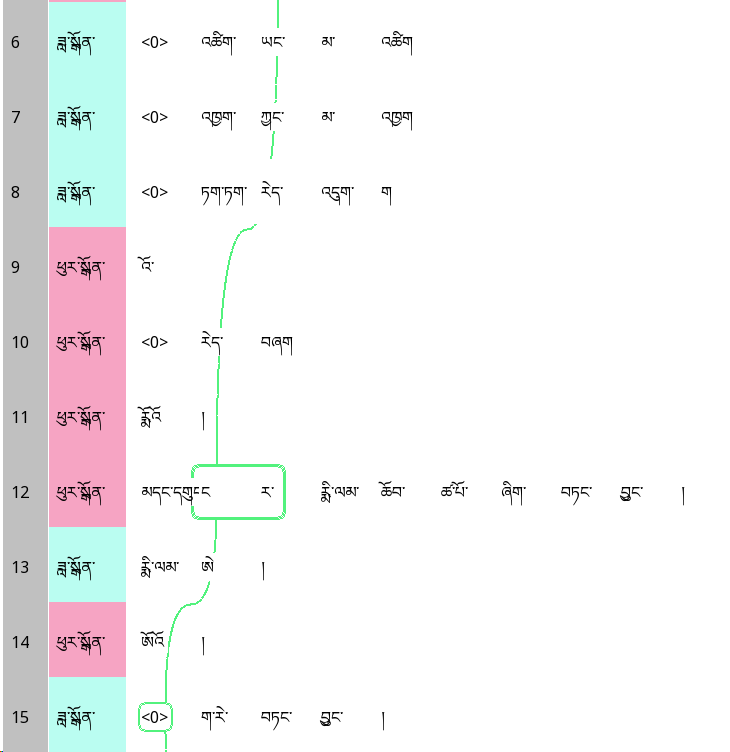
Line 5 is glossed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | བྱས་ཙང་ | ང་ཡང་ | གནམ་ལ་ | བལྟས་སྡོད་ཡོད |
|  | byas.tsang | nga=yang | gnam=la | bltas-sdod-yod |
|  | then | I=also | sky=LOC | look-CONT-EGO |
|  | then | I too | at the sky | am looking |
|  | Then I'm looking at the sky too.' | | | |

The first element, represented by the green chain in the Rezonator screenshot, has a previous mentions value of 1 because there is a grand total of one referential expression in the previous 10 lines that refers to the same person: the one on line 2. The second element has a previous mentions value of 0 because it has never been mentioned before.

To get the number of following mentions, the next 10 lines, i.e. 6-15, are examined:

[A Funny Dream]



|  |  |  |
| --- | --- | --- |
| 6 | Dadrön: | It's not too hot |
| 7 |  | Not too cold |
| 8 |  | Just right, right? |
| 9 | Pudrön: | Yeah // It is. |
| 10 |  | Grandma |
| 11 |  | Last night I had a funny dream. |
| 12 | Dadrön: | A dream? |
| 13 | Pudrön: | Yeah. |
| 14 | Dadrön: | What kind? |

Since there is only one following mention, i.e. ‘I’ on Line 12, the number of following mentions is 1. (The <0> on Line 15 does not count; it is a following zero, covered in the following section.)

Generic and impersonal references are not considered coreferent with any specific entities. For example:

[A baby is coming]

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 35 | Friend: | If the belly is upright, the child is a boy. |
| 36 |  | If the belly is horizontal, |
| 37 |  | well, |
| 38 |  | they say (it) is a girl. |

Here, the child on line 35 is a generic mention. It is therefore not considered coreferent with the child of the pregnant woman being discussed in the lines before line 35, i.e. ‘the child’ on line 35 is a first mention.

However, two generic referential forms are considered coreferent if they are clearly speaking about referring to the same category of entities or the same person in a scenario. For example:

[A Baby is Coming]

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 247 | Friend: | It also creates wounds on the child’s buttocks. |
| 248 | Pregnant woman: | Right. OK. |
| 249 | Friend: | Yeah. |
| 250 | Pregnant woman | Um |
| 251 |  | After the child is uh born, |

Both lines 247 and 251 talk generically about newborn babies. So they are considered coreferential. This is in contrast with the OntoNotes guidelines, where such coreference is only annotated when the subsequent mentions are pronominal.

If a participant first talks about something generically, then talk about a concrete situation that could fit into the generic one, the generic and specific objects are **not** considered coreferent.

## Previous and next zero anaphors (noPrevZero, noNextZero)

‘Zeroes’, i.e. implied participants, are annotated in the text. This is because an entity with high attention may be simply implied in the text without the necessity of mentioning it; hence it is important to include such implications as well.

The identification of zeroes is based entirely on semanto-pragmatics, not on syntax. This has two major consequences:

1. Only **specific** zeroes are annotated; in cases where a missing argument cannot be linked to any mentioned entity in the context (e.g. clauses that may translate to ‘one can …’), no zero is added.
2. Zeroes are retained in subordinate clauses even if they cannot be explicitly stated.

The first decision is reflected in the following example:

[The Lucky Dream]

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 73 | Pudrön | A dictionary and stuff. |
| 74 | Dadrön | A dictionary and stuff. |
| 75 |  | How is (it)? |
| 76 | Pudrön | (It)’s good. |
| 77 |  | (They) sell (it) in the stores. |

In line 77, the dictionary (*it* in the English translation) is a zero because there’s a specific dictionary they’re talking about in the context, i.e. the ones mentioned in the first two lines of the extract. I don’t 90 the seller (*they* in the English translation) a zero since there’s no implied seller in context.

Both of these decisions are reflected in the two following lines.

[The Lucky Dream]

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 88 | Dadrön: | It's just past four. |
| 89 | Pudrön: | There's no point in (us) staying here now. |
| 90 |  | Shall we go <0> drink orange juice? |

Here, the matrix clause is the one with the verb ‘go’, and the subordinate clause is the one with the verb ‘drink’. Both have the same ‘doer’ (or traditionally the ‘subject’), i.e. the two speakers. Although it is not possible to put an explicit agent in the subordinate clause, it is nevertheless shown as a zero. On the other hand, the destination where they are going to get orange juice is **not** marked a zero because there’s no specific destination that they’re talking about from the context.

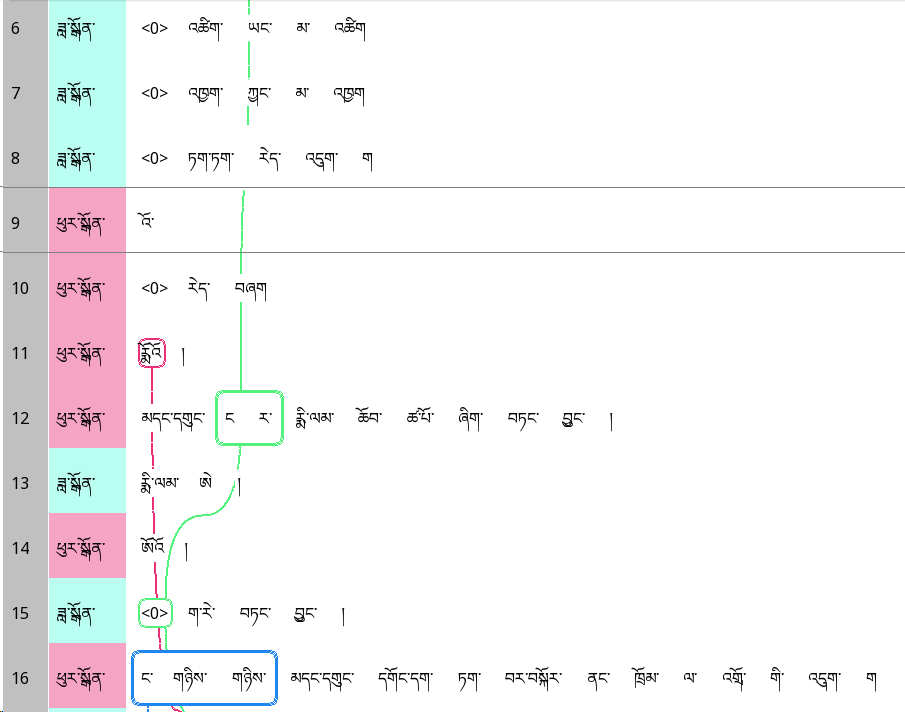
## Bridging / Frame activation (bridge)

The bridge variable is used when something has not been explicitly mentioned or implied by the verb in the previous stretch of discourse, but something highly relevant to it has been mentioned. The main cases of bridging/frame activation annotated are:

* The preceding referent is a group that contains the current referent. (**group-individual**) Note that this does not include very general groups that includes everyone in a certain category (e.g. *men* or *society* with *John*).
* The preceding referent is a member of the group currently being mentioned, **and** the group has been previously established in the conversation, or in general knowledge (e.g. ‘we’ as in the interlocutors of the conversation). (**individual-group**)
* Similar to previous, but **superset-subset** or **subset-superset**
* The current referent is an inherent part of the preceding referent. (**whole-part**). Spatial relationships are classified this way, e.g. top of the couch is part of the couch, the couch is part of near the couch. The time that something happened is considered an inherent part of an event, if framed as such (e.g. *Oh I’ve done that. At* ***that time****, it was a lot easier.*)

These are, to my knowledge, the clearest and least controversial examples; I am conservative in annotating bridging. Consider the following example. The key referential expression I am discussing is ‘the two of us’ (blue box) on line 16. This includes the two speakers, Dadrön and Pudrön. Dadrön has one mention on Line 11; Pudrön has one mention on line 12 and one zero on line 15:

[The Lucky Dream]



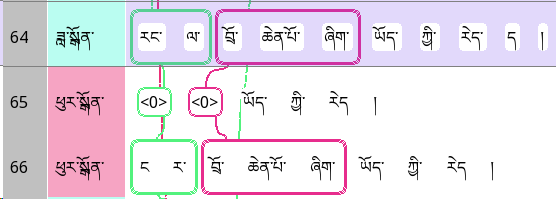
|  |  |  |
| --- | --- | --- |
| 6 | Dadrön: | It's not too hot |
| 7 |  | Not too cold |
| 8 |  | Just right, right? |
| 9 | Pudrön: | Yeah // It is. |
| 10 |  | Grandma. |
| 11 |  | [2] Last night to me a funny dream did. |
|  |  | Last night I had a funny dream. |
| 13 | Dadrön: | A dream? |
| 14 | Pudrön: | Yeah. |
| 15 | Dadrön: | What kind? |
| 16 | Pudrön: | The two of us are going to the Barkor market right. |

Since there was a mention or implication of the two members of the group, I consider the bridge variable to be TRUE. Note that even if only one member of the group were mentioned before; thus, my requirement is weaker than the notion of split antecedence often used in computational linguistics.

## Position in previous multi-argument clauses (justFirst, justLast)

These are continuous variables. justFirst encodes the number of times something has just been the **first** element of a multi-argument clause; justLast encodes the number of times something has just been the **final** element, both in a window of 10 preceding intonation units.

[The Lucky Dream]



|  |  |  |
| --- | --- | --- |
| 64 | Dadrön: | To you good luck may be. |
| 65 | Pudrön: | (I) may have (good luck). |
| 66 |  | I may have good luck. |

Here, on line 66, ‘to me’ and ‘good luck’ have just been the first and last elements on Line 64. Neither referent has been mentioned before Line 64. Hence, the justFirst and justLast annotations are as follows:

|  |  |  |
| --- | --- | --- |
| Element | justFirst | justLast |
| I | 1 | 0 |
| Good luck | 0 | 1 |

## Locality (local)

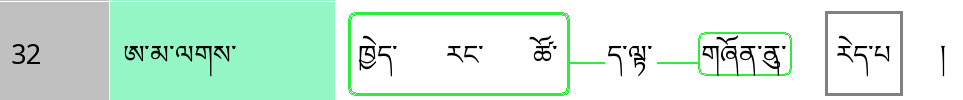
Locality refers to whether an entity can be located by the interlocutors in the immediate spatiotemporal context of the communicative event. This includes:

* Words like འདི་ *here* and ད་ལྟ་ *now* that refer to the current spatiotemporal framework
* Objects in the immediate physical environment. Things like ‘my heart’ are included if talking about a currently valid sentiment.
* ‘I’, ‘you’, and other people who are present.

This does NOT include:

* First or second-person plural pronouns (i.e. ‘we’ or ‘you’) when the expressions include people who are not present. As a result, locality is *not* fully predictable from first and second person. In the following, *khyed.rang.tsho* ‘you’ refers to Dadrön (the addressee) and Dawa (who is not present), and is therefore marked local = FALSE.

[Dadrön and her mother]

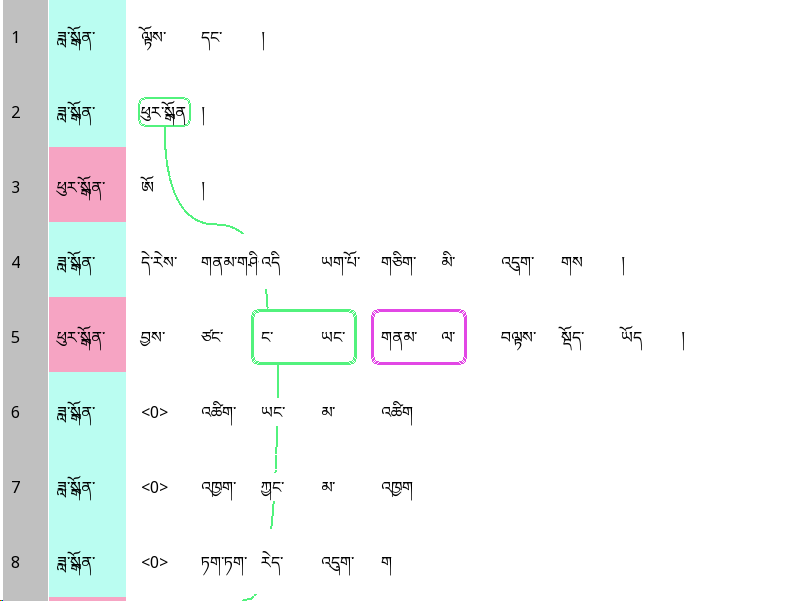


|  |  |  |
| --- | --- | --- |
| 32 | Mother: | You’re still young ones right. |

* Abstract concepts (e.g. I’m **30**, this costs **five dollars**).
* Non-identifiable subsets of local objects (e.g. I want **a kilogram of these potatoes**) (since it cannot be located even if it’s local). The potatoes as a whole would count as local.
* Internal organs unless there’s some recognition of it, e.g. by pointing. While physically they are near, it’s not clear that they are psychologically so. This also goes with other things inside the body, e.g. blood, bile.

Take the following example:

[The Lucky Dream]



|  |  |  |
| --- | --- | --- |
| 5 | Pudrön: | Then I'm looking at the sky too. |

Here, ‘I’ and ‘the sky’ are both local.

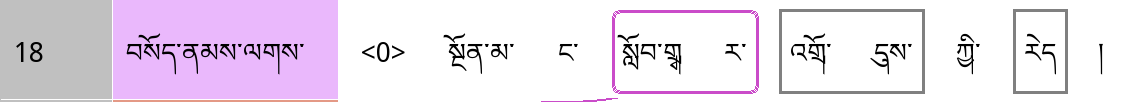
In reported speech, the context in which the original speech event being quoted, not the current speech event, is considered.

## Identifiability (identifiable)

Identifiability is defined as follows: A referent is **identifiable** if the speaker believes that **after listening to the entire utterance**, the listener will be able to connect the referent to an entity whose identity the speaker already know. A referent that is non-identifiable will instead require the listener to create a new identity for referent as the listener is not expected to be able to connect it to an identity they already know.

Identifiability is determined from the point of view of the interlocutors, rather than us. For example:

[The Untrustworthy Neighbour]



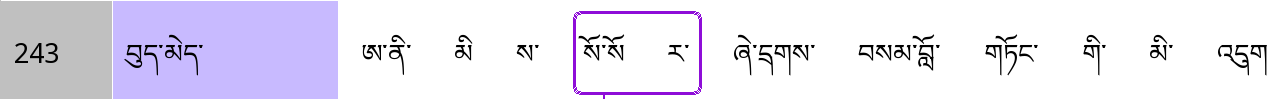
|  |  |  |
| --- | --- | --- |
| 18 | Sonam: | From before when we went to school. |

Though the school is not identifiable for us, it is presumably identifiable for the participants, who know what school they went to together.

Most of the time, identifiability is very clear. Some general principles followed are:

* Generics are considered identifiable, as the listener does not need to create a new ‘file’ for it – it is something already known.
* Elements containing **question words** (who, what, which, etc.) are always coded as non-identifiable. Personal pronouns are generally coded as non-identifiable.
* Referential expressions with words like the indefinite article *zhig* and quantifiers like *kha.shas* ‘some’ are non-identifiable.
* Most of the time, referential expressions with **demonstratives** (other than similative demonstratives) are identifiable.
* In cases of cataphora, the **cataphor** is coded as identifiable as long as the postcedent is identifiable, as the cataphor projects the postcedent.
* If the reflexive/emphatic pronoun སོ་སོ *so.so* is mentioned for the first time, it is non-identifiable since the listener has to create a ‘file’ for this person. Anaphoric སོ་སོ is always coded as identifiable, regardless of who it refers to. For example:

[Two friends reminisce]



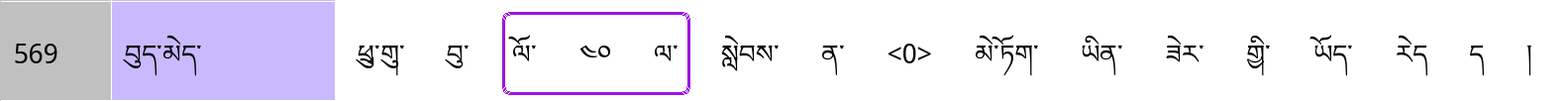
|  |  |  |
| --- | --- | --- |
| 243 | Woman: | Um peoplei don’t give much thought to oneselfj. |

In this case, *oneself* counts as non-identifiable because this is the first mention of the ‘oneself’.

‘ Every’ / ‘ each’ -type NPs with སོ་སོ་ / རེ་རེ་ are treated the same way.

* Specific dates (e.g. last February), times (12.30 today) and ages (e.g. I’m getting close to **30**) are coded as identifiable.

[Two friends reminisce]



|  |  |  |
| --- | --- | --- |
| 569 | Woman: | It is said that men are flowers at age 40. |

* Non-referential arguments, though not technically relevant to identifiability, are coded as non-identifiable.
* Generics are coded as identifiable, since they refer to a class already known, and do not require the listener to open a new ‘file’.
* Propositions are considered non-identifiable unless it has been mentioned before.
* Prices in general are treated as non-identifiable unless mentioned before.

[The Untrustworthy Neighbour]

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| 226 | Yangki: | I paid around 560 for my plane ticket. |

# Role of argument in the clause

## Interrogativity

An element is encoded as Interrogativity = TRUE if it contains a question word like སུ་ *su* who, ག་རེ་ *ga.re* what, ག་གི་ *ga.gi* which, etc. Otherwise it is coded Interrogativity = FALSE.

## Role in the event

The following are roles are recognised in the annotations:

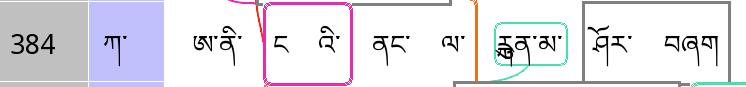
* S – intransitive subject
* A – transitive subject (baseline)
* P – monotransitive object
* T – ditransitive theme
* R – ditransitive recipient
* E – oblique
* COPS – copula ‘subject’
* PRED – copula complement
* EXIST – existent entity in existential clause
* LOC – location in existential clause

Some common decisions on broad classes of verbs are as follows:

|  |  |
| --- | --- |
| Verbs of transfer (saying, etc.) | * Giver – A * Recipient – R * Object transferred – P |
| Benefactive verb | * Dative argument – A * Absolutive argument – P |
| Verb of perception | **Regardless of case**:   * Perceiver – A * Stimulus – P |
| Existential clause | * Location – COPLOC * Existent – EXIST |
| Verbs of movement | * Moving entity – S * Destination, source – E |
| Equative copular clause | * Predicate – PRED * Copula subject – COPS |
| Verbs that take clausal complements | * Speaker/thinker etc. – A * Complement clause – P |
| Analytic causative construction with *bcug* | * Causer – A * Causee – P |

When there are unusual argument structures that do not fall neatly into these traditional categories, the roles were defined with respect to the semantically closest argument structure. For example, in the following example:

[The New Script]



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ཨ་ནི་ | ང | འི་ | ནང་ | ལ་ | རྐུན་མ་ | ཤོར་བཞག |
| a.ni | nga | ='i | nang | =la | rkun.ma | shor-bzhag |
| um | 1sg | =GEN | inside | =LOC | thief | let.loose-PERF.INFER |
|  | COPLOC | | | | EXIST |  |
| 'Um a thief was let loose in my house.' | | | |  |  |  |

I considered *nga’i nang la* ‘in my house’ to be COPLOC and *rkun.ma* ‘thief’ to be EXIST, as the clause primarily serves to convey that there had been a thief in the house.

# Semantic properties of referent

## Animacy

Referential expressions are coded as Animacy = TRUE when it is animate and Animacy = FALSE when it is not.

* All sentient beings, including deities (Buddhas, bodhisattvas, deva, asura) and inferior beings (animals, preta, hell beings) are considered animate.
* Plants are not considered animate.
* Organisations are not considered animate

## Person

This is decomposed into two elements: self (first person) and addressee (second person).

When something is inside a quote, the person of the speaker of that quote, not the actual speaker, is counted as first person.

Impersonals (for example, with སོ་སོ་) are considered third person.

# Formal properties of referential expression

## Length

Length is counted in syllables, measured by the number of *tsheg* (syllable separators in Tibetan writing) + 1. When there are multiple tshegs serving as ellipses, those are counted as a single tsheg only.

## Pronominality

Forms consisting of **only** pronouns (personal, interrogative, indefinite, demonstrative) are coded with pronominality = 1, otherwise pronominality = 0.

མི་ when it means ‘others’ is marked as non-pronominal. Quantifiers are not considered pronominal.