A Syntactically Annotated Corpus of Tibetan

Andreas Wagner and Bettina Zeisler

SFB 441 Universität Tübingen
Nauklerstr. 35, D-72074 Tübingen, Germany
wagner@sfs.uni-tuebingen.de zeis@uni-tuebingen.de

Abstract

This paper describes the creation of a syntactically annotated Tibetan corpus. This corpus forms a part of the TUSNELDA collection of corpora and databases for linguistic research. It will ultimately comprise spoken and written Tibetan texts originating from different regions and historical epochs. These texts are annotated with several kinds of linguistic information, in particular POS tags, phrases, argument structures of verbs, clauses and sentences, as well as several kinds of discourse units and textual segments. The annotation is done in XML. The primary research interest which guides the development of the corpus is the investigation of cross-clausal references, especially the relation between empty arguments (i.e. arguments not overtly realised in a clause) and their antecedents in previous clauses. For this purpose, such references are explicitly encoded so that they can be qualitatively and quantitatively evaluated with the help of standard XML techniques such as XPath search and XSLT transformations. Apart from this primary research interest, we expect that our corpus will be useful for other projects concerning Tibetan and related languages. Like other data in TUSNELDA, it will be made accessible via a WWW query interface.

1. Introduction

This paper deals with the creation of a syntactically annotated corpus of Tibetan. This corpus forms a component of the TUSNELDA data collection, which is being compiled at the Sonderforschungsbereich (special research program) 441 at the University of Tübingen. TUSNELDA (TUebinger Sammlung Nutzbarer Empirischer Linguistischer Datenstrukturen = Tübingen collection of reusable, empirical, linguistic data structures) is a collection of annotated corpora and linguistic databases which cover various linguistic phenomena in a number of languages. Each of these corpora or databases, respectively, is primarily created to serve as an empirical foundation for investigating specific research issues in a particular language or language family.

The primary research purpose guiding the development of the Tibetan corpus can be sketched as follows: Tibetan speakers (and authors) avoid to express information that is given, i.e. that can be derived by the hearer (or reader) from the context. The verb is the only obligatory constituent of a clause, while all nominal constituents can be deleted. In particular, there is no syntactic restriction for the deletion of arguments. Nevertheless, we expect different frequencies of deletion according to the saliency of an argument (corresponding to its semantic role and its position in the verb frame). The investigation of this phenomenon aims at the formulation of rules for the identification of antecedents of empty arguments (i.e. realisations of empty arguments in some previous clause) for different Tibetan varieties. It will also lead to a refined understanding of case relations and semantic roles in Tibetan.

The Tibetan corpus is intended to provide an empirical basis for this research. This corpus will comprise written and spoken texts from different periods: Old Tibetan (8th - 10th century), Classical Tibetan (11th - 19th century), and contemporary West Tibetan, spoken in Ladakh (India) and Baltistan (Pakistan). At present, a Ladakhi "reference text" of 583 clauses has been fully annotated. During the process of annotating this text, numerous issues concerning the information to be encoded, the concrete design of the annotation scheme, and the optimal utilisation of the employed annotation tool had to be solved. The resulting guidelines concerning both the annotation scheme and the annotation procedure have reached a sufficient level of stability to facilitate the efficient annotation of further texts. The corpus will ultimately comprise at least one text from each of the above-mentioned regions and periods.

In the first place, the annotation serves the objectives of the project and reflects the actual working hypotheses. However, we expect that the encoded information will be useful for other research projects concerning Tibetan and other Tibeto-Burman languages as well. To our knowledge, there is currently no other Tibetan corpus which is syntactically annotated.

2. Annotation of the corpus

All texts in the corpus are annotated in XML. The annotation, which is done semi-automatically, provides rich syntactic information about phrasal and argument structures (including semantic roles), as well as information about textual structures.

We decided to encode this information as embedded annotation (i.e. the markup is placed locally at or around the corresponding text) rather than standoff annotation (where the markup is stored in a separate file, including pointers to the primary text). Standoff annotation would be necessary if the structures to be encoded formed overlapping hierarchies, which cannot be modelled in a single XML document. Actually, this problem does not arise for our data. The basic unit to be annotated is a clause, which contains a verb as well as associated arguments and adverbials. The internal structure of a clause can straightforwardly be annotated as an XML hierarchy tree. Cross-clausal references are represented by equally valued ID and IDREF attributes of the corresponding elements. A sentence encompasses a sequence of clauses. Higher textual units (e.g. divisions) essentially consist of a number of sentences. Hence, sub-clausal and super-clausal hierarchies do not overlap so that both can be captured within a single document hierarchy. Concurrent hierarchical units occur only marginally and are not of primary importance. These units concern the physical structure of the annotated texts, e.g. page boundaries; such boundaries are marked by empty XML elements (e.g. <pb/> for a page break), which do not violate the well-formedness of the document.

2.1 Levels of annotation

The lowest level of annotation marks the tokens (i.e. "words") of a text (<tok>) with their orthographic or phonemic realisation (<orth>) and part-of-speech classification (<pos>). To this end, a tagset comprising about 70 POS tags has been devised. The phrase level is encoded by <ntNode> (non-terminal node) elements. An <ntNode> spans an inflectional group within a clause, i.e. a noun phrase if this group forms an argument and an adverbial phrase otherwise. This distinction is marked by an element <ntNodeCat>, which contains the category NP or AvP, respectively. A clause (<clause>) encompasses a verb (always at final position), associated arguments or adverbials, and, if present, embedded clauses. Participle clauses may also be part of an <ntNode>. An element <clauseCat> specifies the type of the clause (simple, chained, embedded, etc.). Tokens, phrases, and clauses may receive a further linguistic description (<desc>). For sub-clausal phrases, this description specifies the case. For verb tokens, the corresponding argument structure is encoded (see below). Above the clause level, sentences (<s>) are marked. In contrast to European languages, a sentence cannot be defined by punctuation. Therefore, we apply the following definition: A typical sentence is a unit of one or more clauses that contains exactly one finite or, alternatively, one non-finite speech-introducing verb. As in any natural language there may be also non-typical sentences, consisting of single words or exclamations, and sentences that are interrupted or not finished. The annotation of the textual level specifies discourse units such as direct or indirect speech, poems or songs, and text segments.

2.2 Argument structure and cross-clausal reference

As mentioned in section 1, the primary research goal underlying the creation of the Tibetan corpus is to examine the relationship between empty (not realised) arguments and their antecedents. The antecedent of an empty argument can be found quite often in the immediately preceding clause (25% of all references) and with decreasing frequency within a distance of 10 clauses (cf. figure 1 below). However, the distance might reach 50 or more clauses, particularly in songs or poems, when similar phrases describing not too complex situations are repeated. Furthermore, chains of corresponding empty arguments which have the same realised antecedent are common. In such a chain, the underlying role and thus the case assignment of the empty argument may vary, e.g. the explicit agent argument of a transitive verb (ergative) might become the sole argument of an intransitive verb (absolutive) or a possessor (aesthetive) in the next clause(s), a patient (absolutive) might become a recipient or beneficient (dative/locative) or even, in rare cases, a transitive agent (ergative) etc. The annotation of argument structure and cross-clausal reference facilitates (a) the quantitative evaluation of the distances between empty arguments and their antecedents and (b) the extraction of chains of empty arguments.

To capture all the information necessary for these purposes, the argument structure of the verbs is modelled by the annotation scheme in the following way: Each verb token receives a serial ID number as an attribute and a special description of its subcategorisation frame (within <desc>). This description comprises (a) the "canonical" argument structure as listed in the lexicon (a list of <complement> elements within a <frame> element), and (b) the "real" frame, i.e. the realisation of the arguments in the clause, including additional arguments (a list of <realComplement> elements within a <realFrame> element). For each canonical and real complement, the semantic role is specified, using an inventory of about 35 roles developed within the project.[1] Furthermore, each canonical complement receives a specification of its case, as does each real complement whose case deviates from the canonical assignment. (Within the <frame> specification, 8 functional case variables are distinguished.[2]) To encode cross-clausal reference, each <realComplement> receives an ID based on the verb number. Empty arguments receive an attribute marking emptiness and a pointer to the antecedent in the text, which in most cases will be a <realComplement> specified in the argument structure of some previous clause. In general, a pointer is encoded as a reference tag (<ref>) with an attribute 'target' that points to the ID number of the corresponding referee.

2.3 Example

An annotation example (a sentence) is given below:

khra·phru·gu *med·tshug* |
child-Abs NEG-exist
'Ø [=They] had no children.'

```
<s>
  <clause>
    <ntNode>
      <tok>
        <orth>khra·phru·gu</orth>
        <pos>NOM:anim~pers</pos>
      </tok>
      <ntNodeCat>NP</ntNodeCat>
      <desc>
        <case>Abs</case>
      </desc>
    </ntNode>
    <tok id="v6">
      <orth n="2">med-tshug</orth>
      <pos>VFIN</pos>
```

1 These roles are differentiated according to their syntactic-semantic functions, such as undergoer, experiencer, agent, patient, etc. as well as with respect to valency and the specific position in the frame.

2 Classical Tibetan has 9 morphological cases: instrumental, dative/locative, locative/purposive, locative, ablative I and II, comitative, genitive, plus absolutive, Ladakhi 6: instrumental, dative/locative, ablative, comitative, genitive, plus absolutive. While only the dative/locative can be used for an experiencer subject or a recipient, we cannot predict which of the locative morphemes will be used for more peripheral arguments, such as beneficiary or direction, nor can we predict when a corresponding postposition is used instead. Likewise we cannot predict which of the ablative morphemes will be used or whether it will not be replaced by a corresponding postposition. For this reason, directional morphemes and postpositions are treated as belonging to abstract case variables. Once the corpus is annotated, the more detailed classification of morphemes and postpositions in the description tag of nominal phrases will lead to a re-evaluation of the verb frames. Additionally we distinguish between the subject case markers ergative (=instrumental) and aesthetive (=dative/locative) and markers for more peripheral arguments, as the subject case markers follow different patterns of case variation.

```
      <desc>
        <feature type="part">NEG</feature>
        <frame>
          <complement>
            <role>POSS</role>
            <case>Aes</case>
          </complement>
          <complement>
            <role>EXST2</role>
            <case>Abs</case>
          </complement>
        </frame>
        <realFrame>
          <realComplement id="v6c1" status="empty">
            <role>POSS</role>
            <ref target="v5c1"></ref>
          </realComplement>
          <realComplement id="v6c2">
            <role>EXST2</role>
          </realComplement>
        </realFrame>
      </desc>
    </tok>
    <clauseCat>simple</clauseCat>
  </clause>
  <punct>|</punct>
</s>
```

The ID of the verb token (id="v6") indicates that it belongs to the 6th clause in the text. Here, the verb canonically selects two arguments, one with *aesthetive* case for the *possessor* and one with *absolutive* case for the *existent* entity possessed. Only the second argument is realised in the clause (by the <ntNode> with absolutive case). The first one is an empty argument whose realised counterpart can be found one clause before. Therefore, there is a pointer from the first <realComplement> (id="v6c1") to its antecedent, the first <realComplement> of the previous clause (id="v5c1").

2.4 Semi-automatic annotation process

For the annotation of the corpus, we use the CLaRK system (Simov et al., 2001). This software comprises a convenient XML editor as well as manifold tools for the automation of certain annotation steps, consistency checking, and searching and extracting information from the annotated documents. These tools can be configured with regard to the employed annotation scheme.

One of these tools is the grammar engine. This facility allows the user to define regular grammars which add some specified XML annotation to sequences of text and/or XML markup which match a certain regular expression. We utilise this mechanism for a semi-automatic annotation at the token level, including POS tagging and the assignment of additional linguistic descriptions. We achieve this by an incremental approach. The texts are sliced into small segments, containing about 100 verbs. After the first slice has been annotated manually, the information encoded for the tokens is transformed into regular grammars, separately for verbs and non-verbs. This transformation is executed by an XSLT processor, which also is part of the CLaRK system. These grammars are applied to the second slice of the text, assigning each token which has already occurred in the first slice the tags <tok>, <orth>, and <pos> together with the corresponding POS tag. The grammar for the verbs also assigns a description <desc> containing the argument frame <frame> and the scheme for the realised arguments <realFrame>, as well as additional features such as negation, question markers, modal auxiliaries, etc. The tokens not captured so far also receive <tok>, <orth>, and <pos> tags, but the correct POS values and descriptions have to be annotated manually. After this has been done, the grammars are rebuilt employing the information about the tokens in the first and the second slice. These updated grammars are in turn applied to the third text slice and so on.

The problem of polysemy of tokens plays a major role mainly within the class of verbs rather than across different parts of speech. Conflicts of POS assignment with respect to nouns and verbs are accounted for by running the corresponding grammars in succession so that ambiguous tokens are processed by the first applied grammar. This yields a small percentage of errors, which are easily detected, since the clause structure is manually annotated. Verbs with more than one reading need a particular set of grammars which require that the annotator manually selects the appropriate reading in the context. Depending on this choice, these grammars assign the <tok> tag with the correct <frame> and <realFrame> structure.

At present the grammars are based on the complete tokens, including all morphemes, which means that each lexeme may have several entries in the grammar; for Ladakhi verbs these may be as many as 30. For the time being, this procedure is the most economic one. However, we plan to develop grammars that can operate on a list of bare lexemes (i.e. the lexicon) and rules for recognising morphemes and their combinations.

Depending on the text type, the three grammars for non-verbs, simple verbs, and polysemous verbs can assign correct POS tags for about 30-50% of the text already after the first turn of annotation, and this rate increases with each subsequent annotation turn. In this way, with increasing coverage of the document (and the corpus), more and more of the annotation can be performed automatically.

A further tool in the CLaRK system which is helpful to facilitate annotation are the so-called value constraints. It is possible to define constraints on the value of certain elements or attributes, depending on their context. The constraint tool can be used to insert values according to such constraints. We employ this mechanism to automatically assign IDs to verb tokens and real complements: The verb token in the n^{th} clause receives the ID "v*n*", and the i^{th} complement of this verb the ID "v*n*c*i*". This assignment is performed by applying an appropriate value constraint.

Another useful application of value constraints concerns the assignment of reference targets. For example, for encoding the pointer from an empty argument to its antecedent, it is possible to define a constraint that restricts the value of the 'target' attribute of the respective <ref> element to the IDs of real complements within the previous 7 clauses. When applied to insert the attribute together with the appropriate value, this constraint generates a menu for choosing among those IDs that meet the restriction. This allows a convenient selection of the appropriate target. (Overriding this constraint is possible to account for those cases where the antecedent's distance exceeds 7.)

Furthermore, additional grammars and constraints help to minimize the amount of time needed for annotation. One grammar, for example, replaces simple abbreviations with complex XML structures, namely the description for the verb token including <frame> and <realFrame>.

3. Evaluation of references

As noted in section 1, the motivation of building the Tibetan corpus is the evaluation of cross-clausal references, in particular the distance between reference and referee and the possible shifts in roles and case marking within chains of reference. Part of the information which is necessary for such an evaluation (e.g. distances) is encoded only implicitly. To make it easily accessible, it has to be represented in an explicit way. For this purpose, XSLT transformations are employed to extract the required information and store it in the annotated document. In detail, the following transformations are performed:

(a) For each empty argument referring to some antecedent, a list of the (possibly multiple) antecedents is extracted and stored within an <antecedList> element in the empty <realComplement> element, including the reference distance (in terms of clauses) as well as the role and the case of each antecedent. Role and/or case mismatches between antecedents and referring arguments are marked. In our example, the distance is "1", role and case are matching.

```
<realComplement id="v6c1" status="empty">
  <role>POSS</role>
  <antecedList>
    <antecedent>
      <ref target="v5c1"></ref>
      <dist>1</dist>
      <role>POSS</role>
      <case>Aes</case>
    </antecedent>
  </antecedList>
</realComplement>
```

This transformation allows collecting statistics of reference distances and role/case mismatches. E.g. in our text of 583 clauses, 406 of 1292 complements are deleted; 272 of these are referential (41 to more than one antecedent). Figure 1 shows the frequencies of reference distances:

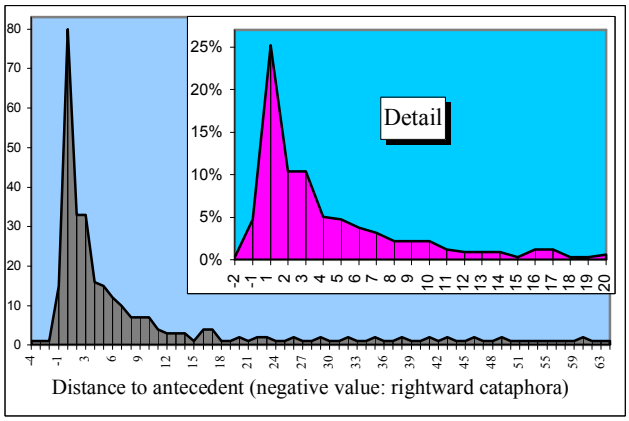

Figure 1: Frequency and distance of references

(b) For each antecedent, a list of referring empty arguments including their distance, their role and case, and an indication of role/case mismatch is extracted and stored within the <realComplement> element representing the antecedent. The element <anaphList> comprising this information receives an attribute 'size' which specifies the number of elements referring to the particular antecedent. This transformation allows the analysis of reference chains and the comparison with other types of anaphora.

```
<realComplement id="v5c1">
  <role>POSS</role>
  <anaphList size="1">
    <anaphor type="empty">
      <ref target="v6c1"></ref>
      <dist>1</dist>
      <role>POSS</role>
      <case>Aes</case>
    </anaphor>
  </anaphList>
</realComplement>
```

4. Next steps

4.1 Accompanying lexical resources

Along with the annotation, we are creating a text-specific lexicon which will list all the lexemes of the particular text. These lexica will be made available with the annotated texts. Verbs will be listed with their (up to four different) stem forms and their frames. A morphological index will be provided for Old and Classical Tibetan and another one for Ladakhi.

Additionally, a complete dictionary of valency is being compiled from extensive fieldwork on Ladakhi verbs. It contains about 900 entries, each comprising the spoken and the written stem forms, the subcategorisation in terms of valency and frame plus an additional semantic subclassification, the translation or meaning of the verb, the Classical Tibetan correspondence and at least one example sentence with interlinear gloss and translation. Examples and spoken forms are further specified with respect to dialect and informant. A main entry may contain several entries for different readings, and within a reading or entry one will also find frames that are derived by means of valency reduction or valency increase.

4.2 Linking and aligning of resources

The different resources mentioned above will be linked to each other in different ways: Each token in the annotated texts will be linked to the corresponding entry in the text-specific lexicon. The corresponding entries of all lexica will be linked. Furthermore, the verb entries will be specifically linked to a concordance file containing the canonical stem forms. This allows searching for variations in the use of stem forms. In addition, an English translation will be aligned to each text by linking the corresponding verbs in the original text and the translation.

4.3 Public access

Like (most of) the other data in the TUSNELDA collection, the annotated texts and the lexica will be made accessible via the WWW. The data will be searchable via a query engine based on XPath and XQuery. A comfortable interface will allow users to formulate search queries and choose among different schemes of output display.

References

Simov, K. & Peev, Z. & Kouylekov, M. & Simov, A. & Dimitrov, M. & Kiryakov, A. (2001). CLaRK - an XML-based System for Corpora Development. In Proceedings of the Corpus Linguistics 2001 Conference pp. 553--560). Lancester, UK: UCREL Technical Papers.