Proceedings of the 9th Seminar of the International Association for Tibetan Studies, Leiden, The Netherlands, June 24-30, 2000 — Information Technology Panel

Automatic Segmentation and Part-Of-Speech Tagging For Tibetan: A First Step Towards Machine Translation

Paul G. Hackett
University of Maryland at College Park
College Park, MD

Abstract

This paper presents what we believe to be the first reported work on Tibetan machine translation (MT). Of the three conceptually distinct components of a MT system — analysis, transfer, and generation — the first phase, consisting of POS tagging has been successfully completed. The combination POS tagger / word-segmenter was manually constructed as a rule-based multi-tagger relying on the Wilson formulation of Tibetan grammar. Partial parsing was also performed in combination with POS-tag sequence disambiguation. The component was evaluated at the task of document indexing for Information Retrieval (IR). Preliminary analysis indicated slightly better (though statistically comparable) performance to n-gram based approaches at a known-item IR task. Although segmentation is application specific, error analysis placed segmentation accuracy at 99%; the accuracy of the POS tagger is also estimated at 99% based on IR error analysis and random sampling.

Introduction

Over the past few decades, great advances have been made in the field of Natural Language Processing (NLP). Much of this research has focused on dominant world languages such as English, French, German, Spanish. In recent years, Asian languages such as Japanese and Chinese have been added to that list, spurred in part by the increased ease with which text in electronic form can be generated, archived and shared, and by the important role those languages play in international commerce. There has been limited research however, for less economically important languages such as Tibetan. This paper presents what we believe to be the first reported work on Tibetan machine translation (MT).

In the construction of a generic machine translation (MT) system, there are three conceptually distinct components: analysis of the source data, transfer of the data to the target mapping, and generation of the target data in appropriate surface forms and configurations. The analysis phase is constituted by Part-Of-Speech (POS) tagging and parsing as the process of taking a sequence of surface forms and converting them into a meaning-preserving internal representation with added lexical information. When performed in a Tibetan context, there is a prerequisite stage of word-

segmentation and sentence boundary detection. We feel that this first phase has been successfully completed.

The subsequent sections of this paper present the theoretical background to the research, an overview of the algorithm and its implementation, issues in evaluation, and a discussion of the current applications of this component.

Background

In a theoretical context, the analysis phase of machine translation can be seen as a set of conceptually distinct processes:

- Segmentation: the separation of a possibly insufficiently differentiated stream of linguistic data into meaningfully quantized groups;
- Tagging: the assignment of one or more part-of-speech tags to each word;
- Parsing: the identification of phrase markers and assignment of a coherent sentence structure to a given set and sequence of tags.

In practical application however, morphological analysis and segmentation are aspects of a single process, and hence some systems blur these distinctions for the sake of efficiency or an advantage in accuracy.

Segmentation

Segmentation can be performed for a variety of reasons: into words for indexing, into sentences for text summarization, etc. Wu proposed a single general principle for segmentation in his work on Chinese which he dubbed the Monotonicity Principle for Segmentation. Stated simply, "[a] valid basic segmentation unit (segment or token) is a substring that no processing stage after the segmenter needs to decompose."[1] The implication of such a principle is that segmentation should be conservative, not prematurely committing to long segments. Wu posited the definition of a valid segment as a substring that no application would ever need to decompose for any reason, whether structural or statistical. While this principle set an upper limit for general-use segmentation length, Wu also presented a lower length limit principle for application-specific uses: that a segmenter should also find the maximum length segment that does not impair accuracy for that application.

Although a number of researchers have advocated a simple "greedy segmentation" algorithm –– longest substring matching to a dictionary, Wu and Fung identified a danger of greedy segmentation (leading to premature commitment) as being a method open to "crossing-segments" errors arising from ambiguous boundaries.[2] As a means of avoiding this danger, Wu likewise advocated performing segmentation in tandem with application specific processes such as name-entity labeling, POS tagging, translation, etc., rather than as a pre-processing stage. In a similar vein, Maosong et al., reported an increase in accuracy in segmentation when segmentation and part-of-speech (POS) tagging were integrated.[3]

POS Tagging and Parsing

A review of tagging research relevant for Tibetan has been provided in elsewhere. Concerning Tibetan however, in brief, Wilson's presentation of Tibetan grammar is centered around

a detailed syntactic classification of verbs. This classification is based on the principle that the syntax of a clause or sentence is determined by the verb that terminates it. In general, Wilson's verb categories are defined both syntactically and semantically as follows:

Class I
- Syntactic Dimension: Nominative Subjects and Nominative Complements
- Semantic Dimension: Existential and Linking

Class II
- Syntactic Dimension: Nominative Subjects and Locative Qualifiers
- Semantic Dimension: Existence qualified referentially, or by location, time, or disposition

Class III
- Syntactic Dimension: Nominative Subjects and Objective Qualifiers
- Semantic Dimension: Reflexive activities qualified by location or destination, and rhetorical statements with a qualified existential identity

Class IV
- Syntactic Dimension: Nominative Subjects and (Non-la class) Syntactic Qualifier
- Semantic Dimension: Existential verbs indicative of a state of separation, absence, conjunction, or disjunction

Class V
- Syntactic Dimension: Agentive Subjects (i.e., agents & instruments) and Nominative [Direct] Objects
- Semantic Dimension: Actions performed by an agent or instrument on something other than itself

Class VI
- Syntactic Dimension: Agentive Subjects (i.e., agents & instruments) and Objective [Direct] Objects
- Semantic Dimension: Actions performed by an agent or instrument on something other than itself

Class VII
- Syntactic Dimension: Dative Subjects and Nominative Objects
- Semantic Dimension: Indicating need, purpose or potential benefit

Class VIII
- Syntactic Dimension: Locative Subjects and Nominative Objects
- Semantic Dimension: Conveying possession, or attribution

These classes provide the basic, first-order categorization scheme for Tibetan verbs. Subcategorization, which has been shown to improve parsing accuracy when combined with lexicalization, is derivable from a consideration of both the semantic scope of the verb class (as is the case with Class IV verbs) and, more significantly, from statistical patterns of usage, though appears to remain invariant across corpora domains. While these patterns of usage vary not only from verb to verb within each class, with sense and, in the case of Class V, VI, and some VIII verbs, voice.

In English, a number of properties differentiate inchoative (or "non-causative") uses of a verb from causative uses. In Tibetan, this alternation of voice between the causative and inchoative is rendered through either the omission of an explicit agent, or the inclusion of a non-sentient agent (an instrument) sometimes with the agent / object sequence reversed for emphasis. Numerous examples illustrate this alternation.

For example, in its causative alternation the Class V (Agentive-Nominative) verb འཛིན་ as "join" or "hold" occurs in the standard agent-object construction: ཁོང་གིས་དཔེ་ཆ་འཛིན་པ་ "He is holding the book," ཁོང་གིས་གསུངས་པ་རྣམས་བློར་འཛིན་པ་ "He is holding in his mind (i.e., has memorized) the teachings," or དགག་བྱའི་ཚད་ལེགས་པར་མི་འཛིན་པ་ "[He] does not have a good grasp on the measure of the object of negation." In an inchoative formation however, a sentence built around the Class V verb འཛིན་ can place emphasis on the subject of an action through the repositioning of words in an object-instrument construction coupled with the lack of a sentient agent. For example: འཇིག་རྟེན་མིམས་བསྐྱེད་ཀྱིས་འཛིན་པ་ "renunciation is joined by (i.e., with) the generation of the [altruistic] mind." Some more examples of Class V Agentive-Nominative verbs in both constructions are:

༡། དམག་རྣམས་ལྷོར་དུ་སྐྱོད་པ།
༢། ཕྱིའི་ཞེམ་པོ་སྐྱེས་བྱའི་རྒྱུད་ཀྱིས་མ་བསྡུས།
1. [He] added the troop to his retinue
2. External matter is not subsumed by (i.e., included within) the continuum of a being

༡། རྒྱལ་ཚབ་དེས་བོད་ལོ་ཉི་ཤུ་བསྐྱངས།
༢། དད་པས་སྐྱོང་།
1. The regent ruled Tibet for 20 years
2. [I] am sustained by faith

༡། ས་གཞི་གཙང་མར་བསྐྲས།
༢། གཉེན་བྱོགས་དབྱ་ཡིས་བསྐྲས།
1. [They] completely cleared the site
2. The friends were separated by anger

As can be seen, there is also a subtle shift in the sense of the verb with the alternation — a feature observed in other languages as well.

Taking this verb classification scheme, and combining it with Wilson's syllable classification yields a complete grammatical formulation suitable for automatic processing. Moreover, since the Wilson formulation of Tibetan grammar incorporates Tibetan–English transfer rules into its parsing strategy, successful identification of a verb's syntactic class yields sufficient sub-categorization information to perform subject–object identification and shallow prepositional phrase attachment.

The Algorithm and Its Implementation

Although simple segmentation can be achieved through longest substring matching to a dictionary, as noted above, algorithms which combine segmentation with POS tagging have been shown to achieve higher accuracy. Statistical methods for POS tagging require a pre-segmented, pre-tagged training corpus, but a rule-based POS tagger can be constructed manually without a training corpus. Since no pre-tagged training corpus exists for Tibetan, a rule-based tagger for Tibetan was manually constructed for this purpose.

Resources

Three types of evidence were employed in the construction of the tagger/parser: lexical knowledge of the words in a language, an explicit knowledge representation that reflects what is known about the ways in which those words can be combined, and statistical evidence of usage patterns gathered from large text corpora. We assembled this lexical knowledge in two forms: a verb lexicon and a noun/adjective dictionary.

Of the two lexicons created for use by the segmentation algorithm, the first was a verb lexicon, necessary for identifying verb-terminated clauses and sentence boundaries. The primary source for the verb lexicon was an augmented version of Wilson's verb classification tables. This was supplemented with other verb tables and glossaries. The resulting verb lexicon used by the segmenter contained over 4,800 distinct verbs. Syntactic class information was obtained by bootstrapping off of correlations with the traditional verb categories of ཐ་དད་, ཐ་མི་དད་, and རྗེས་མཐུན་པ་:

	Wilson Verb Class	Traditional category
I	Nominative-nominative (linking) Verbs	རྗེས་མཐུན་པ་
II	Nominative-locative Verbs	
2.1	simple verbs of existence	རྗེས་མཐུན་པ་
2.2	verbs of living	ཐ་མི་དད་
2.3	verbs of dependence	ཐ་མི་དད་
2.4	verbs expressing attitudes	ཐ་མི་དད་
III	Nominative-objective Verbs	
3.1	verbs of motion	ཐ་མི་དད་
3.2	nominative action verbs	ཐ་མི་དད་
3.3	rhetorical verbs	ཐ་མི་དད་
IV	Nominative-syntactic Verbs	
4.1	separative verbs	ཐ་མི་དད་
4.2	verbs of absence	ཐ་མི་དད་
4.3	conjunctive verbs	རྗེས་མཐུན་པ་
4.4	disjunctive verbs	ཐ་མི་དད་
V	Agentive-nominative Verbs	ཐ་དད་
VI	Agentive-objective Verbs	ཐ་དད་
VII	Purposive-nominative Verbs of Necessity	ཐ་མི་དད་
VIII	Locative-nominative Verbs	
8.1	verbs of possession	རྗེས་མཐུན་པ་
8.2	attributive usage	ཐ་དད་

Because of divergent criteria used in dictionaries for this assessment, this information was then validated against statistical sampling over a text corpus on a verb-by-verb basis.

The second resource needed was a noun/adjective dictionary, necessary for maximum-length substring matching and suffix stripping. This dictionary was formed by initially combining two electronic dictionaries and three dissertation glossaries. Duplicate entries were collapsed, erroneous entries discarded, and a comparison program was used to compare the resultant dictionary against the verb lexicon and remove collocations containing verbs. The final dictionary thus contained roughly 50,000 entries comprised of only nouns, adjectives, and noun-adjective collocations.

Algorithm

The basic algorithm for the Tibetan POS tagger and segmenter is straightforward and exploits the two features of the Tibetan language which make it amenable to analysis: the existence of phrase delimiters and verb termination of sentences and clauses. The process of word-segmentation for Tibetan can be divided into five successive steps:

- Sentence boundary detection
- Verb and verb-clause identification
- Tagging of case-marking particles
- POS tag sequence disambiguation
- Longest substring matching for undifferentiated substrings

We embedded knowledge of some well-understood characteristics of Tibetan into our segmentation algorithm. Since Tibetan is a verb-final language, only the verb lexicon is used to recognize words at the end of a sentence. Sentence boundaries are not marked as such in written Tibetan, but they can be detected fairly reliably by the presence of a verb or some other standard marker (a rhetorical continuative or an ornamental terminating particle) that immediately precedes a marked phrase boundary. We used this as the starting point for shallow parsing of the input text, exploiting syntactic constraints to guide the segmentation process.

We used maximum-length substring matching against the noun/adjective dictionary as a preference criterion for regions in which shallow parsing failed to produce a unique segmentation. These ambiguous cases could result from overgeneration (regions for which the parser generated multiple analyses) or from undergeneration (regions for which the parser failed to discover any syntactic constraints). We used a simple greedy left-to-right search for these regions, removing the longest substring that appeared in our noun/adjective dictionary. When no substring was discovered, we segmented the leftmost syllable as a word of length one and continued.

The parser hinges on sentence boundary detection and verb identification, as shown in the following pseudo-code:

```
for each set of phrases with valid sentence termination
  for each verb
    mark the verb, auxiliaries and Sanskritic adverbs as a verb phrase
  for each remaining substring of unmarked syllables
    mark all known case-marking syllables
  for each remaining substring of unmarked syllables
    perform maximum substring matching against dictionary
```

The final step has the effect of integrating the greedy longest substring matching into the parser itself.

The situation for verb phrases is somewhat more complex. Many Tibetan nouns can be declined into adverbs, a characteristic that Tibetan shares with English. Such adverbial nouns are separated from the verbs they qualify by our parser, so they are segmented as separate words. There is, however, a second class of Tibetan adverbs that correspond to Sanskrit verbal prefixes: རྣམ་པར་ for *vi-*, མངོན་པར་ for *abhi-*, etc. In those instances, adverbs and verbs are bracketed together by our parser and thus segmented as a single word, reflecting

their Sanskritic origin. The last phrase in the example presented below is an instance of this type of verb phrase.

Complete details of the segmentation and tagging process are provided in Hackett (2000). In this paper we provide a brief example to illustrate the parser's operation on two Tibetan phrases:

ཉེ་བ་ཆག་པས་རྩིབས་རྣམས་བཞིན།

དེ་ཞིན་ཀྱིས་ཀུན་ནས་ཡོངས་སུ་བསྲུང་བར་བྱའོ།

[translated roughly as:

"By the hub being broken, similarly the spokes;
you should completely and thoroughly guard that."]

The parser reads in the first phrase. Since there are no sentence terminating particles, and since neither བཞིན་ nor རྣམས་བཞིན་ are verbs, the second phrase is also read into the working buffer. Since the ornamental sentence terminating particle འོ་ is found embedded onto the auxiliary verb ཨ་ the two phrases are successfully matched as a sentence, and the entire last syllable is marked as the closing portion of a verb phrase. Pattern matching is applied to the string of syllables preceding this, identifying the main verb བསྲུང་ "to guard; protect." That verb is followed by a syntactic particle that connects it to the auxiliary verb and preceded by two Sanskritic adverbs, ཀུན་ and ཡོངས་ which are each followed by their respective case-marking particles ནས་ and སུ་. At this point, the entire verb phrase is marked as a unit. The remainder of the sentence is then searched for additional verb clauses. The verb ཆག་ "to break" is found and marked as a verb phrase with no auxiliaries or adverbs, although the verb phrase does contain an additional syllable rendering it as the gerundive ཆག་པ་ with an embedded instrumental case-marking particle ས་. The entire prepositional phrase is then grouped together and marked.

All remaining lexical, syntactic and case marking particles are then found and marked: the plural lexical particle རྣམས་, the frozen adverbial phrase consisting of the single syllable adverb བཞིན་ "similar", the pronoun དེ་, and the single syllable agentive/instrumental case-marking particle ཀྱིས་. Maximum length string matching is then performed for the remaining untagged syllable sequences: ཉེ་བ་, རྩིབས་, and ཞིང་. The words ཉེ་བ་ "hub, center", རྩིབས་ "wheel spoke", and ཞིང་ "you" are found and marked as multi-syllable noun phrases and simple nouns.

The verb lexicon yields the syntactic categorization information for the verb བསྲུང་ (Class V) and the parser successfully identifies both the agent and the nominative object of the verb, with the remainder parsed as a prepositional phrase. The result of the analysis, is the sentence parse-tree (Fig. 1) where, NP indicates a noun phrase, VP a verb phrase, PP a prepositional phrase, INS an instrumental clause, ADV an adverb, and PLP a plural lexical particle.

Evaluation and Implications

The word-segmenter component was evaluated at the task of document indexing for Information Retrieval (IR). Although segmentation is application specific, preliminary analysis indicated slightly better (though statistically comparable) performance to n-gram based approaches at a known-item IR task, with error analysis placing segmentation accuracy at

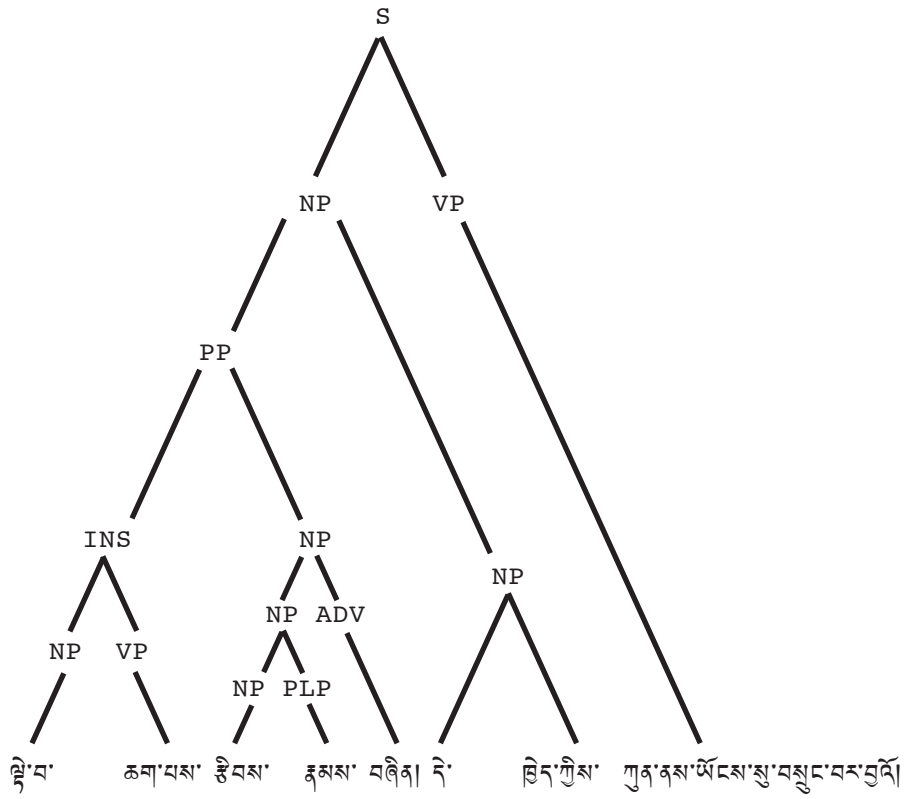

Fig. 1. Sample sentence parse-tree resulting from shallow-parsing based on POS-tagging

99%. The accuracy of the POS tagger is also estimated at 99% based on IR error analysis and random sampling.

The two dominant sources of error for both POS tagging and segmentation were identified during analysis as unknown words (i.e., not contained in the lexicon) and ambiguous boundaries. In the case of unknown words, nearly all instances observed concerned either transliterated Sanskrit mantras or proper names. While some researchers have attempted to compensate for this problem in other languages through the use of a proper name database, it is unclear whether or not this approach would be satisfactory for Tibetan given that proper names have considerable overlap with content-bearing words and phrases. With regard to Sanskrit mantras, however, the problem appears worse. Given the plethora of different mantras, variations in transliteration, and variability in length and composition, it remains unclear if even an explicit list of mantras would suffice. Although an alternate approach consisting of the algorithmic identification of morphologically invalid Tibetan (i.e., Sanskritic) syllables would identify some syllables, any approach short of full lexicalization could not sufficiently identify entire phrases given that some Sanskritic syllables are also valid Tibetan ones. For example, even the simple mantra ཨོཾ་ག་ཏེ་ག་ཏེ་པ་ར་ག་ཏེ་པ་ར་སཾ་ག་ཏེ་བོ་དྷི་སྭཱ་ཧཱ། contains the valid syllables ག་, ཏེ་, བོ་, and ར་. Without the meaning-based determination that, for instance, ཏེ་ is not functioning as a rhetorical continuative, it would be very difficult to automatically parse a sentence containing this phrase correctly, let alone correctly index its constituent words.

The issue of ambiguous boundaries reflects the slightly different difficulty of establishing an unbiased baseline assessment of Tibetan sentences against which the output of any automatic parser could be judged. An example illustrating this point is the following two lines drawn from Śāntideva's *Bodhicaryāvatāra*:

།ཀཻཥ་ཀ་ཡང་ནི་འདོད་པའི་དབེན།

།འཕས་བྱའི་ོོན་ད་ས་དཕྱད་ཕྱིས།

The point of ambiguity in the valid parse of this sentence is whether the word break in the last line is between །འཕས་བྱའི་ོོན་ད་ and ས་དཕྱད་ཕྱིས།, or between །འཕས་བྱའི་ོོན་ད་ས་ and དཕྱད་ཕྱིས།. It is interesting to note that between English translators of this stanza there is a divergence of opinion, however the Tibetan and Mongolian commentators who have written annotations (*mchan*) to the text, notably Thog-med-dpal-bzang-po, Agwangdampa, Kun-bzang-chos-grags, and Gzhan-phan-chos-kyi-snang-ba uniformly take the former interpretation. Similarly, the Sanskrit of this verse accords with the former reading as well, and could be taken as a basis for judgment. This example demonstrates however, the difficulty of finding standards for the resolution of such grammatical ambiguities in cases for which there is no Sanskrit original or divergent opinion on the rendering of earlier compositions, above and beyond the time commitment required to resolve each ambiguity. Consequently, until such fully parsed test collections are compiled, evaluations of computational Tibetan utilities must remain approximate at best.

Current Applications

Despite the work which remains to bring the project to completion, the presently completed component has a number of viable applications. Utilization of the POS / segmentation utility allows for:

- Intelligent full-text indexing and searching
- Cross-language search and retrieval
- Semantic mapping of Tibetan corpora through the clustering of index terms in an n-dimensional vocabulary space ("Latent Semantic Indexing")
- Vocabulary analysis
- Intelligent spell-checking

The first of these points was explored in detail in Hackett.[4] The second application is a natural extension of monolingual indexing. By employing simple word-substitution from a dictionary, a corpus can be indexed for search and retrieval in a separate language. This process has been utilized for numerous parallel text and MT generated text corpora, allowing users to issue queries in a familiar language and retrieve documents in a different language.

The third application has implications for revising contemporary library classification schemes for Tibetan materials. Through the generation of statistically derived index terms for a sufficiently large corpus of Tibetan materials, a set of keywords could be compiled which was immune from any bias or pre-conception on the part of human indexers, in essence, allowing the texts to "speak for themselves."

The fourth and fifth of these uses points to the utility's application to individual texts rather than to a corpus as a whole. By segmenting an individual text, a complete vocabulary list may be obtained with minimal effort. The utility in this is demonstrated not only for simple pedagogical and translation uses, but also in authorial profiling and verification as previously attempted by Valby.[5] Another use is in refining Tibetan spell-checkers. When used, there are two types of spell-checkers currently employed — morphological rules, and high frequency occurrence syllable lists. Employing the form of segmentation described here, single syllable segments (identified as such during segmentation due to typographic error) could be compared in context using longest-substring routines with fuzzy-matching criteria against a lexicon. This would allow for a form of "intelligent" spell-checking and could be used in cases where a corrupt source text is suspected.

Conclusion

We are optimistic that future developments are feasible in terms of both the compilation of necessary resources and the implementation of the relevant aspects of MT theory. Moreover, while the working target language of this project is English, the modular design of the project would allow the utility to be re-targeted for any other language given an appropriate generation module. Although there is no projected timetable for completion, we feel that approximate machine translation for Tibetan to English is attainable within the next few years and will alleviate much of the tedious groundwork in translation, quickly providing researchers with a large corpus of first-pass machine translated texts.

Notes

[1] Wu (1998).
[2] Wu and Fung (1994).
[3] Maosong, et al. (1997).
[4] Hackett (2000).
[5] Valby (1983).

References

ACIP (Asian Classics Input Project), *tshig mdzod chen mo*. (1994) limited distribution. Electronic (Tibetan only) edition; orig. publ.: *bod rgya tshig mdzod chen mo [Great Tibetan-Chinese Dictionary].* 3 vols. Beijing: Nationalities Publishing House (*mi rigs dpe skrun khang*), 1984; reprinted in 2 vols., 1993.

Agwangdampa (*ngag dbang bstan pa*, 1814-1885), *byang chub sems dpa'i spyod pa la 'jug pa'i mchan bu tshig gsal me long*. Reproduced in Choi. Lubsanjab (ed.), *WORKS OF AGWANG-DAMBA (ṄAG-DBAṄ-BSTAN-PA).* New Delhi: Jayyed Pr. (1980), v.1, fol. 1-149a.

Carroll, John, Guido Minnen, and Ted Briscoe, "Can Subcategorisation Probabilities Help a Statistical Parser?" in *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, Montreal (1998).

Chen, Keh-Jiann, and Shing-Huan Liu, "Word Identification for Mandarin Chinese Sentences," *Proceedings of COLING-92*, pp.101-107.

Das, Sarat Chandra, *An Introduction to the Grammar of the Tibetan Language*. Darjeeling (1915); Delhi: Motilal Banarsidass (1972).

Dreyfus, Georges, "Ontology, Philosophy of Language, and Epistemology in Buddhist Tradition." Ph.D. dissertation. Dept. Religious Studies. University of Virginia (1991).

Gzhan-phan-chos-kyi-snang-ba, gzhen-dga' (1871-1927), *byang chub sems dpa' spyod pq la 'jug pa zhes bya ba'i mchan 'grel*. Reproduced in *The Thirteen Great Treatises of Mkhan-po Gźan phan Chos-kyi Snaṅ-ba*, vol. 2, pp. 1-475, Delhi: Jayeed Press, (1987).

Hackett, Paul G., "Approaches to Tibetan Information Retrieval: Segmentation vs. n-grams." Master's Thesis. College of Library and Information Services, University of Maryland, College Park (2000).

Hopkins, Jeffrey, et al., *Tibetan-Sanskrit-English Dictionary*. unpublished.

Kharto, Dorje Wangchuk, *Thumi: dGongs gTer (The Complete Tibetan Verb Forms).* Delhi: Lakshmi Printing Works. n.d.

Kun-bzang-chos-grags (1862?-ca.1940), *byang chub sems dpa'i spyod pa la 'jug pa'i tshig 'grel 'jam dbyang bla ma'i zhal lung bdud rtsi'i theg pa*, in *Collected works gsuṅ bum of Mkhan-po Kun-bzaṅ-dpal-ldan*. Paro, Bhutan: Lama Ngodrub, (1982) v.1 pp.1-741.

Kunsang, Erik Pema, *Concise Dharma Dictionary.* (electronic; work in progress, 1996).

Kwok, K.L., "TREC-5 English and Chinese Retrieval Experiments using PIRCS," *Proceedings of the Fifth Text Retrieval Conference (TREC-5)* Gaithersburg, Md, (November 1997), pp.133-142.

Levinson, Jules Brooks, "The Metaphors of Liberation." Ph.D. dissertation. Dept. Religious Studies. University of Virginia (1994).

Maosong, Sun, Shen Dayang, and Huang Changning, "CSeg & Tag 1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts," *Fifth Conference on Applied Natural Language Processing (ANLP-97)*, pp.119-126 (1997).

Nguyen, Van Be Hai, Ross Wilkinson, and Justin Zobel, "Cross-language Retrieval in English and Vietnamese," *Proceedings of the AAAI-97 Spring Symposium on CLIR*. 1997.

Radford, Andrew, *Syntactic Theory and the Structure of English.* Cambridge: Cambridge University Press (1997).

Roland, Douglas, et al., "Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: The Role of Verb Sense" `http://www.colorado.edu/linguistics/jurafsky/ACL-COMPCORP2000.pdf`.

Śāntideva, *Bodhicaryāvatāra*. Sanskrit published in Leh, Ladakh: Kendrīya Bauddha Vidyā-Saṃsthāna (1989). Tibetan in *byang chub sems dpa'i spyod pa la 'jug pa.* Toh.3871; P.5272.

Thogs-med-dpal-bzang-po (1295-1369), *byang chub sems dpa'i spyod pa la 'jug pa'i 'grel legs par bshad pa'i rgya mtsho*. Thimbu: kun bzang thob rgyal (1975); Sarnath: Pleasure of Elegant Sayings Pr. (1975).

Valby, James M., "The Life and Ideas of the 8th Century A.D. Indian Buddhist Mystic Vimalamitra: A Computer-Assisted Approach to Tibetan Texts." Ph.d. Thesis. Dept. of Far Eastern Studies, University of Saskatchewan (1983).

Wilson, Joe, *Translating Buddhism from Tibetan*. Ithaca: Snow Lion Publ. (1992, 1998).

Wu, Dekai, "A Position Statement on Chinese Segmentation" paper presented at Chinese Language Processing Workshop, Univ.Penn. 30-June to 2-July 1998.

Wu, Dekai, and Pascale Fung, "Improving Chinese Tokenization with Linguistic Filters on Statistical Lexical Acquisition," ANLP-94, Fourth Conference on Applied Natural Language Processing, Stuttgart.

Zahler, Leah Judith, "The Concentrations and Formless Absorptions in Mahayana Buddhism." Ph.D. dissertation. Dept. Religious Studies. University of Virginia (1994).