Abstract:
This introduction surveys research on Tibetan NLP, both in China and in the West, as well as contextualizing the articles contained in the special issue.

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

*Introduction: Tibetan natural language processing**

Nathan W. Hill
SOAS, University of London

Di Jiang
Chinese Academy of Social Science

The current special issue of *Himalayan Linguistics* devoted to Tibetan Natural Language Processing (NLP) emerges from a research project "Tibetan in Digital Communication" (2012-2015) funded by the UK's Arts and Humanities Research Council and hosted at SOAS, University of London. The goals of this project included the creation of a diachronic part-of-speech (POS) tagged digital Tibetan corpus, a Tibetan word segmenter, and a POS-tagger. During the project it was clear that research in the relevant areas was far more advanced in China than in Europe and North America. In addition, in part because of its Chinese language medium and in part because it is formulated in the technical context of engineering, this research was largely unknown and inaccessible to those Anglophone Tibetologists and linguists who stood to benefit enormously from its fruits.

To facilitate the diffusion of this Chinese learning, with the kind sponsorship of the Sino-British Fellowship Trust, our research project hosted Professor Jiang Di, co-editor of this special issues. Professor Jiang stayed in London for nearly two weeks, participated in a range of meetings, and offered a well-attended talk on the achievements of Tibetan NLP research in China.[1] In the course of his visit we agreed on this special issue as a follow-up endeavor. Professor Jiang and his colleagues have shown the utmost patience and fortitude in the execution of this project. We also thank Kristine Hildebrandt, Andrew Hardie, You-Jing Lin, and the editorial board of *Himalayan Linguistics* for their cooperation. The reader will understand the challenges of crossing the divides of national academic cultures and disciplinary orientations and we hope she will be inspired to take up similar burdens, charting the vistas opened here. Nonetheless, sobriety compels us to fear that the benefits of NLP, both social and academic, will not be immediately clear to the reader. Consequently, this introduction transitions away from autobiography and toward history. We offer a brief general contextualization of NLP, its components and its achievements. We then present a history of Tibetan NLP, including both research in China and, more briefly, research outside of China. We end with a few observations on the contents of the current issue.

* This work is supported by National Natural Science Foundation of China (No. 60173024).

[1] We here thank the Chinese Embassy for gracing his talk and the following reception with a delegation.

1 Natural language processing: The general context

To a profound extent NLP technologies underpin the life of all citizens of developed nations. Technologies from the iPhone to Google Translate build upon decades of meticulous research by linguists and engineers alike. In the late 1990s, the NLP research community developed the notion of a Basic Language Resource Kit (BLARK). BLARKs consist of sets of computational tools foundational to the creation of language technologies, from spellcheckers to Apple's Siri. BLARKs also enable cross-linguistic communication, including machine-aided translation and inter-lingual software for the Internet. A hierarchical organization of necessary components to a BLARK includes the following:

1. Script is available in Unicode
2. Digital texts are available in quantity
3. Segmenter: a software tool that divides tokens
4. Tagger: a software tool that assigns part-of-speech categories to each token
5. Lemmatizer: a software tool that associates different morphological forms of the same word (i.e. different tokens with different POS-tags) under one lemma
6. Parser: a software tool that models higher order syntactic analysis

This is a non-exhaustive list, with many other potential BLARK components, such as machine readable dictionaries or statistical language models, implied along the way.

In the better studied languages NLP has served not only commercial and practical technological and communicative goals but has also shaped and facilitated the trajectory of ivory-tower research. One may name the many TreeBank corpora to have emerged from the Pennsylvania school of historical syntacticians as but one example. Any field linguist who has toiled for months over the transcription and glossing of fieldwork data, longs to let this cup pass to a homunculus of boundless energy and patience. With an annotated corpus one can find in seconds the crucial examples to answer a linguistic question. In the days of pen and ink accumulating the same evidence was a lifetime's labour.

2 The current status of Tibetan language processing

Although Tibetan NLP research lags behind research on the commercially paramount languages such as English and Chinese, a considerable amount of work is available and progress is being made quickly. Given that the majority of Tibetan speakers live in China, it is no surprise that Chinese research leads the way in Tibetan NLP.

2.1 Research in the People's Republic of China

In the last 20 years, eight domains of Tibetan language information processing have evolved in China: (1) encoding standards, (2) resource collection and statistics, (3) machine readable dictionary compilation, (4) dictionary sorting and retrieval, (5) Tibetan OCR and speech recognition, (6) word segmenting and POS-tagging, (7) syntactic and semantic analysis, (8) machine translation. Here we give a brief survey of these domains and their developments in China, especially concentrating on text processing and syntactic analysis, which represent the future of Tibetan

information processing. Upon completing this overview, readers still curious about the history of Tibetan NLP research in China may consult two relevant papers of Jiang (2003, 2006).

2.1.1 Encoding and other standards

The development of electronic resources for Tibetan involves both agreement on encoding standards and the creation of input applications. There are two relevant encoding standards, the standard maintained by the International Standard Organization (ISO/IEC) and the Chinese standard (GB). The ISO standard for Tibetan, ISO/IEC 10646-1(1993/P.DAM.6-ISO/IEC JTC1/SC2/ WG2 N 2627:1995), was passed at the Stockholm meeting in 1995 as proposed by the Chinese delegation. Two interacting Chinese government standards, GB16959-1997 and GB/T 16960.1-1997 were publicly announced two years later.

Standards in other areas are still under development. These include the transliteration of Tibetan into Latin characters, word segmentation, part-of-speech tag sets, corpus development, etc. In 2015, the Department of Language Information Management of the Ministry of Education published a green book on standards of Tibetan to Latin transliteration, word segmentation and part-of-speech classification of Tibetan, formulated by the Institute of Ethnology and Anthropology of Chinese Academy of Social Sciences and Minzu University of China.

2.1.2 Resource collection and statistics

In the early 1990s, researchers began to study the statistical profile of Tibetan as revealed in corpora, at the level of akṣaras, syllables, and words. The earliest paper, dealing with the statistical attributes of the basic components of Tibetan writing in relationship to phonological structure, was published in 1992 (Jiang). Another early paper was "a handling to the Tibetan characters of spatial construction as linear sequence" (Jiang and Dong 1994). Using a statistical approach to the frequently used words in modern Tibetan, these investigations obtained data such as the average length and frequency of words (Jiang and Dong 1995). "The Phonological Construction of Tibetan Words and Its Frequency Phenomena", delivered at the international meeting on multi-character processing at Waseda university in 1998 (Jiang 1998) provides further statistical analysis on the structure of Tibetan akṣaras and on the use of letters to compose akṣaras. Statistical research on large-scale historical Tibetan texts has also been undertaken, e.g. "<中华大藏经·丹珠尔>藏文对勘本字频统计分析" (Zhaxiceren 1997), and "Researches of Calculations of Tibetan Characters, Pieces, Syllables, Vocabulary and Universal Frequency and its Applications" (Lu 2003). In order to better understand the properties of Tibetan text, the paper "Principles on the Selection of Tibetan Sample Corpus Text for Computer Statistics" presents a thorough analysis of ancient and modern Tibetan texts. The notable feature of this paper is a classification of Tibetan texts in terms of theme, type of writing and style of writing, as well as the establishment of sampling principles, which lay a solid foundation for research on larger scale Tibetan corpora (Zhou and Jiang 1999).

Based on the statistical analysis of the lexicon of written Tibetan, some scholars have undertaken the statistical analysis of Tibetan texts, and in particular, applying Markov models, they have analyzed information entropy. In "the Entropy of Written Tibetan and related problems", the data of the zero rank entropy, rank one entropy and rank two entropy have been computed (Jiang 1998). Through comparing the figures obtained with those of English, Russian, and Chinese, the information content of different language sources can be diagrammed, which is an aid to language engineering and processing. As far as writing is concerned, the information entropy statistics in

Tibetan consists of four levels: letters, akṣaras, syllables and words. Although the earliest discussions about Tibetan information entropy only treated letters, the conclusions were nonetheless useful.

Compared with other writing systems (such as Chinese), the information entropy of Tibetan akṣaras has its own characteristics. A follow-up computing of the entropy is conducted on the word-forming component or akṣaras of written Tibetan in a large-scale corpus (with approximately 40 million akṣaras). By way of illustration, 786 distinct akṣaras have been extracted from the Tibetan Buddhist Canon, hence the rank one entropy is 4.80, the rank two entropy is 3.12, the rank three entropy is 2.70, and the superfluous degree is 0.72, while the respective corresponding values in English are: rank one entropy: 4.03, rank two entropy: 3.32, rank three entropy: 3.1, superfluous degree: 0.71. From the comparison, one can conclude that the information resource of Tibetan is close to that of English. The statistics of superfluous degree also clearly indicate the attribute of the information resource of written Tibetan, and the superfluous degree of the Tibetan characters shows that the redundancy of Tibetan source text is about 72%, that is to say, when a Tibetan text is composed, 72% of the information is predetermined by the structure of the words and the language (characters, words, and sentences), and only 28% is optional, which means that three quarters of the letters in Tibetan will not convey information but ensure that the combinations of letters will accord with the principles in Tibetan of word formation, character-formation, and grammar concerned (Yan and Jiang 2004).

Admittedly, the statistic data and the entropy computing of the Tibetan texts are not purely theoretical, but more important in terms of language engineering. The post-processing of character recognition in Tibetan that we are undertaking makes full use of the disambiguation function based on entropy, which has effectively improved accuracy rate of identification.

2.1.3 Machine-readable dictionary compilation

A Tibetan machine readable dictionary is a foundation for natural language processing in Tibetan. In natural language processing, the basic vocabulary or morphemes in a language are the most reliable and effective knowledge resources. In terms of the characteristics of Tibetan, a machine-readable dictionary should have its individual property in addition to the properties shared by all machine-readable dictionaries. The foremost task in compiling such a dictionary is to decide on the basic principles of compilation, including the general principle, the principle of standard, the principle of grammar, the principle of stability, word-formation and bilingualism (Jiang 2000). The next task is to provide as exhaustive as possible the lexical description of each individual word (and partial syntactic description) by means of grammatical tagging and labeling, including labeling part-of-speech and pragmatic attributes (honorific vocabulary, Buddhist vocabulary, Sanskrit loans, and variant forms of words). Based on these principles, the department of IEA, through many years' efforts, has compiled a large-scale machine-readable Tibetan dictionary with more than 100,000 entries, each with some grammatical attribute labeling (Jiang 2005). In addition, a large text corpus with syntactic tagging of more than one million akṣaras has been produced (Long 2012; Kang 2013).

2.1.4 Dictionary sorting and retrieval (compositor and index)

In order to solve the practical problems in engineering application and deepen the understanding of the attributes of the Tibetan corpus, scholars in China have conducted successive compositor and index studies of Tibetan. Like studies in all other languages, the text processing and dictionary-merging in Tibetan are also faced with the problem of reordering or reindexing. And the

Tibetan compositor is even more complex, because the coding property involves the complex two-dimensional word-formation in Tibetan. Therefore, it is necessary to determine the Tibetan word sequence first in theory. In terms of principles of word sequence in other languages, the letter writing of European languages such as English may be termed "simple-conventional sequence", where letters are limited in number and arranged linearly, for example "abc" and "level"; Chinese features graphic characters, and the characters are complex, without fixed structural sequence, which may be termed "non-conventional sequence". The structural sequence of Chinese characters may be mathematically analogized to the typical dispersed structure of the Markov space, such as "乙，川，㙓，爨，鸚" (with diversified strokes). In contrast, though Tibetan consists of letters, yet the terrace two dimensional structure possesses the graphic property; due to the property of structural sequence, construction layer and character sequence in its structure, it may be viewed as "complex conventional sequence". As the sort order of a given script system and orthographic practice of a given language are conventionalized by a given ethnic culture and there is no wrong or right in cultural conventions, we must follow the traditional dictionary sequence in dealing with Tibetan electronic texts.

According to this principle, in their paper "On the Sequence of Tibetan Words and the Method of Making Sequence" (Jiang and Zhou 2001), Jiang Di and Zhou Jiwen discussed the basic principles and approaches to Tibetan sorting. The scheme and flow chart employed in this paper is a theoretical foundation for advanced researches of Tibetan language engineering. Moreover, this article has produced new knowledge by making use of experimental methods, and has combined traditional culture with modern science and technology.

In order to realize the basic principle of Tibetan sorting and establish a mathematic model and algorithm for the compositor, a result of an experiment has been worked out called "Mathematic Model and Algorithm of written Tibetan compositor", which was published in 计算机学报 *Chinese Journal of Computers* (Jiang and Kang 2004). In addition, the applied research on the Tibetan texts index has been published in English as well in the journal *Studies in Language and Linguistics* (Kang and Jiang 2004).

2.1.5 Tibetan OCR and speech recognition

Research on optical character recognition (OCR) in the case of Tibetan may be aimed at one of three distinct problems: printed character recognition, handwriting character recognition, and xylograph character recognition. Until now, the handwriting and printed recognition of Tibetan characters has made a great progress, which achieves a practical level, but the recognition research of ancient woodblock documents is still in the study stage. In the 2000s, many papers on Tibetan character recognition were published. In 2012, a number of important papers were anthologized in the book 藏文识别原理与应用 *The principles and application of Tibetan Character recognition* (Jiang, ed. 2012).

There are several institutions in China working on Tibetan speech synthesis and recognition, including the Institute of Ethnology and Anthropology and the Institute of Linguistics of the Chinese Academy of Social Sciences, the Department of Chinese Language and Literature of Peking University, the Institute of National Languages Information Technology of Northwest University for Nationalities, Qinghai Normal University, Tibet University, etc. These organizations mainly concentrate on fundamental research, in particular developing Tibetan acoustic parameter libraries, direct progress on Tibetan speech recognition has not begun.

In recent years, some large companies, such as Xunware Technology Co. Ltd., have also become involved in research on Tibetan speech recognition, and actively participate in the speech recognition research of minority languages such as Mongolian, Tibetan and Uygur. These companies have successively established some speech recognition laboratories of ethnic languages in Tibet University, Qinghai Normal University, China Ethnic Languages Translation Bureau etc., which greatly promoted the development of speech recognition of ethnic languages.

2.1.6 Word segmentation and part-of-speech tagging

Research on Tibetan word segmentation began in China in 1999, when the main method was dictionary matching. Later the scheme of chunk word segmentation was adopted (Jiang 2003), but the results were not good and not scalable. Statistical-based research on Tibetan word segmentation gradually increased around 2010. Hidden Markov Model and Maximum Entropy Model were used in the early days. Since 2011, the method of syllable-based tagging has been introduced into Tibetan word segmentation. In this approach word breaking is treated as a POS-tagging problem over syllables. However, in the process of its introduction, some researchers did not combine the method of syllable-based tagging with the features of Tibetan language, and did not deal well with the abbreviated syllables (e.g. the genitive ཀྱི་), which lack any analog in Chinese. HD Liu, CJ Long, CJ Kang, D Jiang etc. specially set two word positions for Tibetan abbreviated syllables based on their features, and successfully applied syllable-based tagging to Tibetan word segmentation, which achieved significant results. Nevertheless, the problem has not been fully solved.

Research on Tibetan POS-tagging have made some achievements, but there are still many problems: the first is the lack of consistency in the standards of different researchers for POS-tagging; the second is the lack of consistency in the corpus used to train and test POS-tagging; the third is that the different tagging systems have not been disclosed, consequently, it is difficult to make an objective evaluation of various systems. All research to date has adopted a statistical model for POS-tagging, but there are not enough Tibetan texts for statistical training, and a high number of out-of-vocabulary words also affects the tagging accuracy.

2.1.7 Syntactic and semantic analysis

Syntactic and semantic research of Tibetan has been comparatively weak. Integral syntactic analysis is difficult to achieve, and partial syntactic analysis mainly amounts to chunking. Researches such as chunk boundary recognition, chunk type labeling, and semantic role labeling are conducted through the adoption of chunk division systems of different sizes (L Li 2013; TH Wang 2013; CJ Long 2014). Currently, researchers of the Institute of Ethnology and Anthropology of Chinese Academy of Social Sciences are building a treebank of phrase hierarchy structures, and attempting to gradually realize the conversion between a constituent-based treebank and a dependency treebank.

2.1.8 Machine translation

The earliest research in China on machine translation between Tibetan and Chinese began in the 1990s, but little progress was achieved before the Seventh National Workshop on Machine Translation in 2011, when the machine translation evaluation of Tibetan and Chinese was involved for the first time in the machine evaluation organized by the Institute of Computing Technology of Chinese Academy of Sciences, and the translation model was mainly government documents. Since then machine translation of Tibetan and Chinese has become a hot topic of research. There have

been a number of units developing machine-aided translation systems for Tibetan and Chinese. In 2015, the national evaluation result of machine translation system showed that, there were three teams participating in Tibetan-Chinese machine translation, viz. Tibet University, Xiamen University and the Institute of Information Technology of China. Compared with previous years, the test field and translation effects did not change significantly. Up to now, there are three online machine translation systems for Tibetan and Chinese, which are the Sunshine Translation System of Tibet University, the Multilingual Machine Translation System of China Ethnic Languages Translation Bureau, and Torangetek Multilingual Translation System of the Institute of Computing Technology of Chinese Academy of Sciences.

2.2 Tibetan NLP research outside of the PRC

Although there have been repeated efforts in Tibetan NLP research outside of the PRC these efforts generally have not built on each other, but remained isolated interventions. Paul Hackett has worked since the late 1990s on a variety of topics including information retrieval, and POS-taggging (Hackett and Oard 1997, 2000; Oard et al. 1998; Hackett 2000a, 2000b, 2003, 2010, 2013). Unfortunately, his software and corpora have not been made publically accessible. A project on Tibetan zero-anaphora at Tübingen University from 2002-2008 published two programmatic studies (Wagner and Zeisler 2004; Zeisler 2004). The corpus produced by this project is too small to reward reuse in other research. The project 'Tibetan in Digital Communication' has produced a number of research papers, including descriptions of the tag-set (Garrett et al. 2015) and rule based POS-tagger (Garrett et al. 2014; Garrett and Hill 2015a), as well as new linguistic insights the corpus has revealed (Garrett et al. 2013; Garrett and Hill 2015b). The software tools, corpus and workflow system of this project are being made publically available (https://github.com/tibetan-nlp). The SOAS part-of-speech tagger achieves > 99.8% accuracy with an average ambiguity of 1.38 tags per word (as of 6 March 2016).

Quite aside from NLP research per se, projects outside of China have made great strides in the collection of Tibetan electronic corpora. The largest electronic corpus is the ever expanding e-text library of the Tibetan Buddhist Resource Center,[2] which as of December 27, 2014 consisted of 959,020 pages of text. These texts are encoded in Unicode and stored in XML files. The material for this collection comes from two sources: OCRed modern printed texts and the digital files of publishers of Tibetan texts. The TBRC provides a dedicated search interface, but the corpus itself is not available for download. The Old Tibetan Documents Online (OTDO) is a collection of 109 Old Tibetan texts.[3] The texts include documents discovered at the library cave at Dunhuang and imperial inscriptions form central Tibet. These materials are not included in any other digital corpus. OTDO texts are encoded in a purposed designed Latin transcription. The OTDO includes a search interface; the corpus is downloadable. Otani Tibetan E-Texts consists of 14 texts input from xylographs held at the Otani University library.[4] The bulk of this collection is historical and biographical classics. These texts, in Unicode, are available for download. The collection is not searchable online. A digital version of the Derge Kanjur (an edition of the Tibetan Buddhist canon), prepared by the British Library and SOAS, University of London is hosted by the Tibetan and Himalayan Digital Library

[2] http://www.tbrc.org

[3] http: //otdo.aa.tufs.ac.jp/ and http://otdo.aa-ken.jp/

[4] http://web1.otani.ac.jp/cri/twrpw/results/e-texts/

of the University of Virginia.[5] The data are in Unicode and stored in XML. There is a search facility. Unfortunately, the edition currently online contains many typos. The TBRC in collaboration with Eusukhia[6] have proofread these materials, but the corrected version is not yet available for public download or consultation. To date the corpus produced by 'Tibetan in Digital Communication' is the only one available to include POS-tagging.

3 The contents of the current issue

The current issue brings together contributions devoted to a suite of topics in Tibetan NLP by scholars from China, Europe, and North America. At the more brass tacks end of the spectrum Rowinski and Keutzer discuss a system for Tibetan OCR and Ma and Wu treat a system for on-line Tibetan handwriting input. The papers of Jiang and Li and of Meng and Jiang discuss linguistic properties of Tibetan, respectively the classification of verbs and an elaboration of compound adjectives, which become important in the design and implementation of NLP systems. The bulk of the papers seek to enrich lightly annotated corpora (e.g. with part-of-speech tagging) with more sophisticated annotation. Li et al. seek to implement a system for Tibetan functional chunk recognition. Hindt outlines a method for annotating syntactic trees. Long and Li further attempt semantic role labeling. Jia et al. discuss a system for identifying Tibetan personal names. Zhao et al. propose a means of extracting examples of trisyllabic light verb constructions from a corpus in which they are not explicitly annotated. More at the level of practical applications Liu et al. explore the possibilities of Chinese to Tibetan machine aided translation and Schmidt discusses the use of NLP tools and research in language learning. From this selection of papers the reader will achieve a good overview of diversity of Tibetan NLP research going on today. We hope that by making available this work in one place, this issue will also lower the transaction cost for those researchers themselves hoping to make a contribution to Tibetan NLP or who may use similar approaches in the study of other Tibeto-Burman languages.

Acknowledgements

The editors would like to thank Dr. Long CJ and Dr. Liu HD, who give us much help in preparing this paper and assisting in organizing the special issue of the journal. This work is supported by the National Natural Science Foundation of China (61132009, 31271337), the National Social Science Foundation of China (12&ZD174, 10&ZD124), and the UK's Arts and Humanities Research Council.

References

Garrett, Edward; and Hill, Nathan W. 2015a. "A constraint grammar POS-Tagger for Tibetan." In: Proceedings of the Workshop on Constraint Grammar - Methods, Tools and Applications, at NODALIDA 2015, May 11-13, 2015. Vilnius: Institute of the Lithuanian Language, pp. 19-22.

[5] http://www.thlib.org/encyclopedias/literary/canons/kt/catalog.php#cat=d/k

[6] http://www.esukhia.org

Garrett, Edward; Hill, Nathan W.; Kilgarriff, Adam; Vadlapudi, Ravikiran; and Zadoks, Abel. 2015. "The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries." *Revue d'Etudes Tibétaines* 32: 51-86.

Garrett, Edward; and Hill, Nathan W. 2015b. "Constituent order in the Tibetan noun phrase." *SOAS Working Papers in Linguistics* 17: 35-48.

Garrett, Edward; Hill, Nathan W.; and Zadoks, Abel .2014. "A rule-based part-of-speech tagger for Classical Tibetan." *Himalayan Linguistics* 13.1: 9-57.

Garrett, Edward; Hill, Nathan W.; and Zadoks, Abel. 2013. "Disambiguating Tibetan verb stems with matrix verbs in the indirect infinitive construction." *Bulletin of Tibetology* 49.2: 35-44.

Hackett, Paul. 2000a. "Approaches to Tibetan Information Retrieval: Segmentation vs. n-grams". University of Maryland MLS thesis.

Hackett, Paul. 2000b. "Automatic segmentation and part-of-speech tagging for Tibetan". Paper presented at the Ninth Seminar of the International Association for Tibetan Studies (IATS-9), Leiden, The Netherlands, June 2000.

Hackett, Paul. 2003. "An Entropy-based Assessment of the Tibetan Unicode Encoding". Paper presented at the Tenth Seminar of the International Association for Tibetan Studies (IATS-X), Oxford, United Kingdom, September 2003.

Hackett, Paul. 2010. "The use of yig-cha and chos-kyi-rnam-grangs in computing lexical cohesion for Tibetan topic boundary detection". Paper presented at the Twelfth Seminar of the International Association for Tibetan Studies (IATS-XII), Vancouver, British Columbia, August 2010.

Hackett, Paul. 2013. "Digital resources for research and translation of the Tibetan Buddhist Canon". Paper presented at the Thirteenth Seminar of the International Association for Tibetan Studies (IATS-XIII), Ulaanbaatar, Mongolia, July 2013.

Hackett, Paul; and Oard, Doug. 1997. "Document translation for cross-language text retrieval at the University of Maryland". Paper presented at the Sixth Text REtrieval Conference (TREC-6), Gaithersburg MD, November 1997.

Hackett, Paul; and Oard, Doug. 2000. "Comparison of word-based and syllable-based retrieval for Tibetan". Presented as a poster at the Information Retrieval for Asian Languages Workshop in Hong Kong in September, 2000.

Jiang, D. 1992 "Statistics on the inflexional phenomenon of Tibetan verbs". *Minzu Yuwen* 4: 47-50. 江荻（1992）：藏语动词屈折现象的统计分析，《民族语文》，第 4 期。

Jiang, D. 1998. "An Entropy value of Classical Tibetan language and some other questions". In: Chen, LW. (Ed.), *1998 Collections of International Coference for Chinese information processing*, 377-381. Beijing: TsingHua University Press. 江荻（1998）：书面藏语的熵值及相关问题，黄昌宁主编：《1998 年中文信息处理国际会议论文集》，第 377－381 页。北京：清华大学出版社。

Jiang, D. 1999. "The application of concordance technology to Tibetan corpus. In: Huang, CN.; and Dong, ZD. (Eds.), *Collections on computational linguistics*, 359-364. Beijing: TsingHua University Press. 江荻（1999）：语篇索引技术在藏文文本中的应用，黄昌宁，董振东主编：《计算语言学文集》，第 359－364 页。北京：清华大学出版社。

Jiang, D. 2003. "A new perspective for modern Tibetan machine processing and its development: an insight into the method of computerized automatic understanding of natural languages in terms of chunk identification". In: Xu, B.; Sun, MS.; and Jin, GJ. (Eds), *Some important problems in Chinese language information processing*, 438-448. Beijing: Science Press of China.

江荻（2003）：现代藏语的机器处理及发展之路，徐波，孙茂松，靳光瑾主编：《汉语自然语言处理若干重要问题》，第 438-448 页。北京：科学出版社。

Jiang D. 2003a. "On syntactic chunks and formal markers of Tibetan". In: Sun, MS; Chen, QX. (Eds), *Language calculation and content-based text processing*, 160-166. Beijing: Tsinghua University Press. 江荻（2003）：现代藏语的句法组块与形式标记，孙茂松，陈群秀主编：《语言计算与基于内容的文本处理》，第 160-166 页。北京：清华大学出版社。

Jiang, D. 2003b. "The method and process of the definition to syntactic chunks in modern Tibetan". *Minzu Yuwen* 4: 30-39. 江荻（2003）：现代藏语组块分词的方法和过程，《民族语文》，第 4 期，30－39。

Jiang, D. 2003c. "Recognition and information extraction of finite verbs in Modern Tibetan". In: Sun, MS; Yao, TS; and Yuan, CF (Eds), *Advances in computation of Oriental languages*, 154-160. Beijing: Tsinghua University Press. 江荻（2003）：现代藏语谓语动词的识别与信息提取，Maosong Sun, Tian Shunyao, Chunfa Yuan(eds. 2003). *Advances in Computation of Oriental Languages*, Pp154-160. Beijing: Tsinghua University Press.

Jiang, D. 2006. "The history and advance in the text information processing of Tibetan language". In: Cao, YQ.; Sun, MS. (Ed.), *Frontiers of Chinese information processing*, 83-97. Beijing: Tsinghua University Press. 江荻（2006）：藏语文本信息处理的历程与进展，载：曹右琦，孙茂松（主编），中文信息处理前沿进展—中国中文信息学会二十五周年学术会议，第 83-97 页。北京：清华大学出版社。

Jiang, D.; and Dong, YH. 1995. "Research on property of Tibetan characters as information processing". *Journal of Chinese Information Processing* 9.2: 37-44. 江荻，董颖红（1995）：藏文信息处理属性统计研究，《中文信息学报》，第 2 期。

Jiang, D.; and Kang, CJ. 2004a. "The methods of lemmatization of bound case forms in Modern Tibetan". *2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, IEEE Press.

Jiang, D.; and Kang, CJ. 2004b. "The sorting mathematical model and algorithm of Written Tibetan language". *Journal of Computer* 4: 524-529. 江荻，康才畯（2004）：书面藏语排序的数学模型及算法，《计算机学报》，第 4 期 524-529。

Jiang, D.; and Long, CJ. 2003. "The Markers of non-finite VP of Tibetan and its automatic recognizing strategies". In: Sun, MS; Yao, TS; and Yuan, CF (Eds), *Advances in computation of Oriental languages*, 169-175. Beijing: Tsinghua University Press.

Jiang, D.; and Dong, YH. 1994. "A handling to the Tibetan characters of spacial construction as linear order". *Chinese Information Processing* 4: 44-46. 江荻，董颖红（1994）：藏字叠加结构线性处理统计分析，《中文信息》，第 4 期。

Jiang, D.; and Zhou, JW. 2001. "On the sequence of Tibetan words and the method of making sequence". *Journal of Chinese Information Processing* 1: 56-64. 江荻，周季文（2001）：藏语的序性及排序方法，《中文信息学报》第 1 期。

Jiang, D.; Long, CJ.; and Zhang, JC. 2005. "The verbal entries and their description in a grammatical information-dictionary of contemporary Tibetan". In: Dale, Robert; Wong, Kam-Fai; Su, Jian; and Kwong, Oi Yee (Eds), *Natural language processing- IJCNLP2005*, 874-884. Berlin: Springer.

Jiang, Di; and Dong, YH. 1995. "Research on property of Tibetan characters as information processing". *Journal of Chinese Information Processing* 9.2: 37-44. 江荻，董颖红（1995）：藏文信息处理属性统计研究，《中文信息学报》，第 2 期。

Kang, CJ; and Jiang, D. 2004. "The optimized index model of Tibetan dictionary". *Studies in*

Language and Linguistics 1: 120-125.

Kang, CJ.; Long, CJ.; and Jiang, D. 2013. "Tibetan word segmentation based on word-position tagging". Paper presented at 2013 International Conference on Asian Language Processing, IALP. Urumqi, China, Aug 17-19.

Long, CJ. 2012. "Key issues in the text information processing of Tibetan language". *E-science Technology and Application* 3.4: 51–58. 龙从军 2012 藏语文本信息处理的几个关键问题。《科研信息化技术与应用》第 3 卷第 4 期: 51-58。

Lu, YJ.; Ma, SP.; Zhang, M.; and Luo, G. 2003. "Researches of calculations of Tibetan characters, pieces, syllables, vocabulary and universal frequency and its applications". *Journal of Northwest Minorities University* 24.2: 5-55. 卢亚军，马少平，张敏，罗广 2003 基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究。《西北民族大学学报》第 2 期。

Oard, D; Dorr, BJ; Hackett, P; and Katsova M (1998). A Comparative Study of Knowledge-Based Approaches for Cross-Language Information Retrieval. CS-TR-3897, UMIACS Tech. Report Library

Wagner, Andreas; and Zeisler, Bettina. 2004. "A syntactically annotated corpus of Tibetan." *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisboa, May 2004.*

Zeisler, Bettina. 2004. "An annotation of what is not there: Empty arguments and cross-clausal reference in spoken and written Tibetan texts." *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories, Tübingen, December 2004.*

Nathan W. Hill
nh36@soas.ac.uk

Di Jiang
jiangdi@cass.org.cn