The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

Christopher Handy
Michael Litchard

handyca@mcmaster.ca
http://handyc.sdf.org

Inaugural NARNiHS Conference
23 July 2017

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

This study demonstrates a practical method for extracting recurrent strings from digitized texts in cases where grammar, vocabulary and other information about the texts are partly or entirely unknown.

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

Our method involves building concordances of words and phrases from digitized input sets of texts, using a simple but effective pattern-recognition algorithm. The algorithm can be generalized to work with information in any language, but we restrict this study to just three major languages of the Buddhist tradition: Sanskrit, classical Tibetan and classical Chinese.

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

We utilize free text files available in online databases so that our examples can be verified easily. We also provide source code examples of our algorithm in C and Haskell, available on GitHub: https://github.com/handyc

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

Three languages of the Buddhist literary tradition — Sanskrit, Tibetan and Chinese — lack word boundary delimiters in their traditional manuscripts. A human being familiar with these languages can identify individual words in such manuscripts, but the texts have no spaces between words, such that we cannot locate words without prior knowledge of the language.

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

As a result of this problem, digital concordances for Sanskrit, Tibetan and Chinese often require some amount of manual part-of-speech tagging before sending datasets to the computer.

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

Our software runs through every possible string of ‘n’ syllables in every text of a text corpus, where ‘n’ is any number desired by the user. We refer to strings of connected syllables as n-grams, named by their specific ‘n’ size as 1-gram, 2-gram, 3-gram, etc.

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

The program counts each n-gram in a given text, and creates a concordance file for that text. These concordance files are then combined to form a concordance file for the entire corpus.

The Haks Language Modeling System: Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

For example, searching on ‘n’=4 among a sample set of major Sanskrit Mahāyāna texts, we found the sequence bo-dhi-sa-ttva appearing frequently in many different texts. If this sequence appears more frequently in Mahāyāna texts than in mainstream texts, it could be used as an identifier for the Mahāyāna genre.

Sanskrit Mahāyāna and Mūlasarvāstivāda Vinaya frequent n-grams

Mahāyāna 4-gram	Mahāyāna 8-gram	MSV 4-gram	MSV 8-gram
bodhisattva (3343)	nuttarāyāṁsamyaksaṁbo (451)	kathayati (1367)	bhagavataārocaya (166)
tathāgata (2096)	ttarāyāṁsamyaksaṁbodhau (436)	bhagavanā (550)	vaśeṣāmāpattimāpa (166)
kulaputra (1346)	nuttarāṁsamyaksaṁbodhi (320)	gavānāha (520)	ghāvaśeṣāmāpattimā (166)
śatasaha (1298)	samyaksaṁbodhimabhisaṁ (303)	sakathaya (499)	gavataārocayanti (166)
bodhisattvā (1289)	ttarāṁsamyaksaṁbodhima (302)	samayena (458)	nakālenatenasama (164)
samyaksaṁbo (901)	rāṁsamyaksaṁbodhimabhi (297)	kathayanti (405)	kālenatenasamaye (164)
bodhisattvo (874)	thāgatorhansamyaksaṁbu (266)	saṃlakṣaya (372)	lenatenasamayena (163)
athakhalu (860)	kṣetraparamāṇurajaḥ (262)	damavoca (372)	vobhagavataāroca (159)
lokadhātu (812)	ddhakṣetraparamāṇura (260)	lakṣayati (356)	kṣavobhagavataāro (159)
sahasrāṇi (746)	traparamāṇurajaḥsa (259)	bhagavatā (346)	tatprakaraṇaṃbhikṣavo (153)
mahāmate (740)	buddhakṣetraparamāṇu (250)	midamavo (332)	tprakaraṇaṃbhikṣavobha (152)
buddhakṣetra (647)	gavantametadavoca (241)	ghāvaśeṣā (311)	raṇaṃbhikṣavobhagava (152)
tasahasrā (637)	bhagavantametadavo (236)	gṛhapati (303)	karaṇaṃbhikṣavobhaga (152)
dhisattvasya (616)	ṭīnayutaśatasaha (224)	praticchannā (298)	bhikṣavobhagavataā (152)
tadavoca (610)	koṭīnayutaśatasa (224)	tisakatha (292)	ṇaṃbhikṣavobhagavata (151)
sarvasattva (599)	paramāṇurajaḥsamā (205)	tadabhava (288)	gavantamidamavoca (150)
sattvomahā (578)	myaksaṁbodhimabhisaṁbu (205)	rvavadyāva (280)	ārocayantibhagavā (135)
tathāgatā (571)	ṭīniyutaśatasaha (196)	saṃghāvaśe (278)*	tenakālenatenasa (135)

Tibetan *'dul ba* (=*vinaya*, “monastic law”) texts from the Derge *Kanjur* (Buddhist canon)

'dul ba text	2-gram	4-gram
kl00001e1.txt	DGE_SLONG (2671)	SO_SOR_THAR_PA'I (672)
kl00001e2.txt	PA_DANG (1827)	BCOM_LDAN_'DAS_KYIS (578)
kl0001e3inc.txt	DGE_SLONG (2031)	BCOM_LDAN_'DAS_KYIS (523)
kl0001e4inc.txt	PA_DANG (1713)	BCOM_LDAN_'DAS_KYIS (407)
kl00002e1.txt	DGE_SLONG (346)	TSE_DANG_LDAN_PA (101)
kl00003e1.txt	DGE_SLONG (2089)	TSE_DANG_LDAN_PA (647)
kl00003e2inc.txt	DGE_SLONG (2636)	TSE_DANG_LDAN_PA (749)
kl00003e3.txt	DGE_SLONG (2550)	BCOM_LDAN_'DAS_KYIS (879)
kl00003e4.txt	PA_DANG (1679)	BCOM_LDAN_'DAS_KYIS (591)
kl00004e.txt	DGE_SLONG (465)	DGE_SLONG_MA_GANG (211)

Tibetan *'dul ba* (=*vinaya*, “monastic law”) texts from the Derge *Kanjur* (Buddhist canon)

'dul ba text	2-gram	4-gram
KL00001-001(eTB).txt	དགེ་སློང་ (2622)	སོ་སོར་ཐར་པའི་ (667)
KL00001-002(eTB).txt	ཡ་དང་ (1796)	བཅོམ་ལྡན་འདས་ཀྱིས་ (568)
KL00001-003(eTB).txt	དགེ་སློང་ (1993)	བཅོམ་ལྡན་འདས་ཀྱིས་ (515)
KL00001-004(eTB).txt	ཡ་དང་ (1686)	བཅོམ་ལྡན་འདས་ཀྱིས་ (404)
KL00002-001(eTB).txt	དགེ་སློང་ (342)	ཚེ་དང་ལྡན་པ་ (100)
KL00003-001(eTB).txt	དགེ་སློང་ (2042)	ཚེ་དང་ལྡན་པ་ (622)
KL00003-002(eTB).txt	དགེ་སློང་ (2595)	ཚེ་དང་ལྡན་པ་ (733)
KL00003-003(eTB).txt	དགེ་སློང་ (2500)	བཅོམ་ལྡན་འདས་ཀྱིས་ (858)
KL00003-004(eTB).txt	ཡ་དང་ (1656)	བཅོམ་ལྡན་འདས་ཀྱིས་ (581)
KL00004(eTB).txt	དགེ་སློང་ (457)	དགེ་སློང་མ་གང་ (211)

Tibetan *mdo mang* (=*sūtra*) texts from the Derge *Kanjur* (Buddhist canon)

mdo mang text	4-gram	8-gram
kl00094e.txt	BYANG_CHUB_SEMS_DPA' (111)	BYANG_CHUB_SEMS_DPA'_MCHOG_TU_DGA'_BA'I (25)
kl00095e.txt	BYANG_CHUB_SEMS_DPA' (608)	SLONG_DAG_DE_LTAR_BYANG_CHUB_SEMS_DPA' (36)
kl00096e.txt	BYANG_CHUB_SEMS_DPA' (55)	BYANG_CHUB_SEMS_DPA'I_RAB_TU_'BYUNG_BA (25)
kl00097e.txt	BYANG_CHUB_SEMS_DPA' (63)	PA_JI_LTAR_NA_BYANG_CHUB_SEMS_DPA' (30)
kl00098e.txt	BYANG_CHUB_SEMS_DPA' (29)	BYANG_CHUB_SEMS_DPA'_SEMS_DPA'_CHEN_PO (11)
kl00108e.txt	BYANG_CHUB_SEMS_DPA' (187)	BYANG_CHUB_SEMS_DPA'_SEMS_DPA'_CHEN_PO (88)
kl00353.txt	BYANG_CHUB_SEMS_DPA' (353)	BYANG_CHUB_SEMS_DPA'_SEMS_DPA'_CHEN_PO (22)
kl00357.txt	BYANG_CHUB_SEMS_DPA' (27)	BYANG_CHUB_SEMS_DPA'_SEMS_DPA'_CHEN_PO (8)

Tibetan *mdo mang* (=*sūtra*) texts from the Derge *Kanjur* (Buddhist canon)

mdo mang text	4-gram	8-gram
KL00094(eTB).txt	བྱང་ཆུབ་ཏུ་ནི་ (977)*	བྱང་ཆུབ་ཏུ་ནི་སེམས་བསྐྱེད་དོ་བདེ་ (679)*
KL00095(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (591)	སྒོང་དག་དེ་ལྟར་བྱང་ཆུབ་སེམས་དཔའ་ (36)
KL00096(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (55)	བྱང་ཆུབ་སེམས་དཔའི་རབ་ཏུ་འཇུང་བ་ (24)
KL00097(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (60)	པ་ཌི་ལྟར་ན་བྱང་ཆུབ་སེམས་དཔའ་ (29)
KL00098(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (29)	བྱང་ཆུབ་སེམས་དཔའ་སེམས་དཔའ་ཆེན་པོ་ (11)
KL00108(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (108)	བྱང་ཆུབ་སེམས་དཔའ་སེམས་དཔའ་ཆེན་པོ་ (87)
KL00353(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (232)	བྱང་ཆུབ་ཏུ་སེམས་བསྐྱེད་པར་ཐུར་ཀོ་ (9)*
KL00357(eTB).txt	བྱང་ཆུབ་སེམས་དཔའ་ (26)	བྱང་ཆུབ་སེམས་དཔའ་སེམས་དཔའ་ཆེན་པོ་ (8)

Chinese texts from the Taishō canon (Chinese Buddhist canon)

Text	1-gram	2-gram	4-gram	8-gram
T01	不 (21400) 有 (21078) 是 (20109) 者 (16553)	比丘 (7863) 如是 (7848) 世尊 (4876) 沙門 (3014)	所以者何 (879) 沙門瞿曇 (799) 尊者阿難 (557) 爾時世尊 (544)	如來無所著等正覺 (246) 聞佛所說歡喜奉行 (242) 我聞如是一時佛遊 (216) 誦我聞如是一時佛 (199)
T02	是 (21383) 不 (18528) 時 (16702) 如 (16470)	比丘 (11395) 如是 (7862) 世尊 (7260) 爾時 (6083)	爾時世尊 (2592) 所說歡喜 (1803) 聞佛所說 (1756) 佛所說歡 (1699)	衛國祇樹給孤獨園 (1505) 舍衛國祇樹給孤獨 (1505) 聞佛所說歡喜奉行 (1380) 如是我聞一時佛住 (1275)
T03	是 (16212) 如 (15204) 不 (14458) 無 (14028)	菩薩 (5812) 如是 (5750) 爾時 (4446) 一切 (3834)	藐三菩提 (1094) 三藐三菩 (1094) 阿耨多羅 (1093) 耨多羅三 (1093)	阿耨多羅三藐三菩 (1092) 耨多羅三藐三菩提 (1092) 成阿耨多羅三藐三 (349) 多羅三藐三菩提心 (241)
T04	不 (16280) 是 (11786) 者 (11086) 人 (10931)	比丘 (3117) 是故 (2268) 世尊 (2248) 爾時 (2141)	是故說曰 (1052) 爾時世尊 (520) 即說偈言 (347) 亦復如是 (300)	舍衛國祇樹給孤獨 (138) 衛國祇樹給孤獨園 (137) 在舍衛國祇樹給孤 (137) 佛在舍衛國祇樹給 (134)

Chinese texts from the Taishō canon (Chinese Buddhist canon)

Text	1-gram	2-gram	4-gram	8-gram
T08	不 (30632) 菩 (29114) 是 (28812) 無 (27650)	波羅 (16532) 羅蜜 (16492) 菩薩 (16387) 般若 (12207)	般若波羅 (11852) 若波羅蜜 (11851) 菩薩摩訶 (6444) 薩摩訶薩 (6439)	阿耨多羅三藐三菩 (2176) 耨多羅三藐三菩提 (2170) 須菩提菩薩摩訶薩 (878) 阿耨多羅三耶三菩 (878)
T09	一 (19709) 無 (18855) 佛 (16633) 法 (15684)	一切 (14366) 菩薩 (9990) 眾生 (8800) 如來 (4534)	一切眾生 (2769) 菩薩摩訶 (2089) 薩摩訶薩 (2088) 令一切眾 (1187)	阿耨多羅三藐三菩 (544) 耨多羅三藐三菩提 (544) 子是為菩薩摩訶薩 (267) 佛子是為菩薩摩訶 (267)
T10	一 (30535) 無 (29172) 切 (21625) 諸 (20996)	一切 (21590) 菩薩 (17716) 眾生 (11979) 如是 (6543)	一切眾生 (3496) 菩薩摩訶 (2369) 薩摩訶薩 (2368) 一切諸佛 (1211)	阿耨多羅三藐三菩 (495) 耨多羅三藐三菩提 (495) 謂菩薩生如是心我 (316) 薩生如是心我已得 (316)
T11	無 (25990) 如 (21719) 是 (21493) 不 (19870)	菩薩 (10722) 如是 (7850) 一切 (7281) 如來 (6397)	薩摩訶薩 (2303) 菩薩摩訶 (2303) 文殊師利 (1033) 一切眾生 (802)	阿耨多羅三藐三菩 (770) 耨多羅三藐三菩提 (770) 舍利子菩薩摩訶薩 (331) 得阿耨多羅三藐三 (154)

Chinese texts from the Taishō canon (Chinese Buddhist canon)

Text	1-gram	2-gram	4-gram	8-gram
T12	是 (31334) 不 (28306) 如 (27667) 無 (27300)	菩薩 (10028) 如是 (9206) 如來 (7684) 眾生 (7624)	菩薩摩訶 (1900) 薩摩訶薩 (1896) 亦復如是 (1640) 一切眾生 (1385)	阿耨多羅三藐三菩 (1052) 耨多羅三藐三菩提 (1050) 得阿耨多羅三藐三 (369) 善男子菩薩摩訶薩 (334)
T13	無 (23427) 是 (22261) 不 (21834) 如 (19803)	菩薩 (9957) 一切 (9957) 眾生 (7919) 如是 (7743)	菩薩摩訶 (1686) 薩摩訶薩 (1680) 一切眾生 (1109) 阿耨多羅 (905)	阿耨多羅三藐三菩 (898) 耨多羅三藐三菩提 (898) 多羅三藐三菩提心 (263) 發阿耨多羅三藐三 (248)
T14	無 (58875) 佛 (44990) 南 (37115) 如 (19559)	南無 (36603) 佛南 (28455) 如來 (9873) 菩薩 (7861)	如來南無 (4964) 王佛南無 (2403) 佛南無無 (2148) 菩薩南無 (1460)	阿耨多羅三藐三菩 (607) 耨多羅三藐三菩提 (607) 中華電子佛典協會 (332) 曰智慧是為六何謂 (242)
T15	不 (21311) 無 (21014) 是 (19035) 如 (15385)	菩薩 (7368) 一切 (5846) 如是 (4631) 眾生 (3574)	文殊師利 (1282) 不可思議 (1125) 菩薩摩訶 (645) 薩摩訶薩 (644)	阿耨多羅三藐三菩 (359) 耨多羅三藐三菩提 (359) 童子菩薩摩訶薩復 (171) 子菩薩摩訶薩復有 (171)

Chinese texts from the Taishō canon (Chinese Buddhist canon)

Text	1-gram	2-gram	4-gram	8-gram
T21	一 (15785) 二 (15538) 如 (11616) 不 (11430)	二合 (7689) 一切 (5164) 如是 (4365) 菩薩 (3513)	娑嚩二合 (1151) 此陀羅尼 (784) 嚩二合引 (767) 電子佛典 (684)	中華電子佛典協會 (456) 項本資料庫可自由 (228) 電子佛典普及版完 (228) 電子佛典協會版權 (228)
T22	比 (38638) 丘 (37624) 不 (33402) 是 (25496)	比丘 (37618) 諸比 (9580) 丘尼 (8891) 如是 (8760)	若比丘尼 (1663) 六群比丘 (1570) 白佛佛言 (1517) 告諸比丘 (1456)	默然故是事如是持 (509) 從今是戒應如是說 (482) 今是戒應如是說若 (449) 是戒應如是說若比 (442)
T23	不 (30543) 是 (28299) 比 (23647) 丘 (22827)	比丘 (22825) 苾芻 (8356) 如是 (6004) 丘尼 (4509)	種種因緣 (1156) 僧伽婆尸 (1054) 伽婆尸沙 (1054) 白佛佛言 (1011)	應如是說若復苾芻 (409) 學處應如是說若復 (372) 處應如是說若復苾 (368) 制其學處應如是說 (363)
T24	不 (27289) 者 (23089) 是 (18720) 有 (16271)	苾芻 (8486) 比丘 (6489) 世尊 (5131) 如是 (4908)	爾時世尊 (695) 白佛佛言 (611) 時諸苾芻 (601) 波逸底迦 (551)	波逸底迦若復苾芻 (279) 苾芻以緣白佛佛言 (206) 者波逸底迦若復苾 (191) 逸底迦若復苾芻尼 (182)