Sanskrit lexicography, past and future

Gérard Huet

Inria Paris

Abstract

We give a short survey on past achievements of Sanskrit lexicography, and consider new methodologies for the design of modern digital Sanskrit dictionaries.

1 Ancient Sanskrit Lexicography

There have been numerous efforts to develop Sanskrit dictionaries, starting in antique times with the *Nighaṇṭu* of Yāska (7th century B.C.), together with its commentary the *Nirukta*, a glossary of Vedic terms. The purpose of this effort was to preserve the Vedic texts, already hard to decipher by that time.

In the 5th century B.C. the linguist Pāṇini wrote a complete grammatical treatise on the language in 8 parts (*Aṣṭādhyāyī*), that explained precisely how to formally derive both a correct enunciation and a paraphrase of its meaning, according to the locutor's intention. More than a mere grammar of the language, it is actually a non-deterministic algorithm generating linguistic forms as phonemic streams derived from root syllables and meta-linguistic operators (*anubandha*). This grammar, together with its commentaries by the later grammarians Katyāyana and Patañjali, established the standard of correct Sanskrit speech, and thus became *de facto* normative.

The *Aṣṭādhyāyī* theorizes correct utterances as sequences of word forms respecting functional dependencies through a notion of semantic role (*kāraka*). These word forms are glued together through a process of phonetic smoothing (*sandhi*) to produce continuous speech. This linguistic generative process operates on sound, discretized as a set of 50 phonemes (*varṇa*). Writing merely records the resulting phonemic stream, just grouping it in syllables that are encoded as ligatures of basic glyphs for consonants and vowel. This can be done in any Indian syllabic script, but the most usual one is the Northern *Devanāgarī* script. In any case, word boundaries are lost in the *sandhi* phonetic smoothing, and thus, in written text as well as in continuous oral enunciation, it is not possible to directly identify the forms of lexicalized words. Thus reading must proceed by first segmenting enunciations into sequences of word forms (*padapāṭha*), undoing *sandhi*.

Numerous auxiliary treatises are needed to use the *Aṣṭādhyāyī* grammar. First of all, *dhātupāṭhas* are databases of verbal roots, equipped with markers (*anubandhas*) in order to document the stem formation and other morphological parameters. Then a catalogue of families of nominal stems, sharing similar morphological characters, called the *gaṇapātha*. For instance, the list starting with *kiśara* (*kiśarādi*) gives all nominal stems admitting the secondary (*taddhita*) suffix *ṣṭhan* usable for constructing a masculine derivative in *-ika* (resp. feminine in *-ikī*) denoting a merchant of such items, according to Pāṇini's *sūtra* (IV,4,53), such as *kiśarika* (fem. *kiśarikī*) for a merchant of perfumes. Such open-ended lists were augmented now and then, in order to reflect linguistic usage not covered by the grammar. Another important such resource is the *uṇādisūtrakośa*, that gives supplementary rules (*sūtra*) listing primary nominal stems, not available in a direct way as derivatives of verbal roots. Other auxiliary treatises describe gender or accent prescriptions [4].

The grammatical tradition established by Pāṇini was finalized by Patañjali around 150 B.C. in his comprehensive commentary *Mahābhāṣya*. Pāṇini's system became the *de facto* standard of the language, classical Sanskrit, whose usage from thereon evolved only within the limits of the grammar. This makes Sanskrit unique as a human language. Contrary to the standard modern linguistic view, where a language is defined from its corpus of attested sentences, we may consider classical Sanskrit as actually *defined* by Pāṇini's grammar. Indeed, Sanskrit *usage* evolved, and it makes sense to date Sanskrit corpus according to its style, but it kept within the conceptual structure of the *Aṣṭādhyāyī*.

Sanskrit is contrasted with Prakrit, which literally means "natural". The various Prakrits were the vernacular languages of North India in ancient times. For instance, Magadhī was the administrative language of the Maurya Empire at the time of Alexander. This was the language spoken by Buddha Śakyamuni. Prakrits were local variations evolved from the languages spoken in Northern India in Vedic times. But Sanskrit was not the vernacular language of any place or time. It was actually never the maternal language of anyone. It was (and still is) a learned language. Brahmin families would send their sons to the *gurukula* boarding school, around age 8, to be drilled in Sanskrit by rote learning. Since the teaching was based on the grammar, with the teacher requesting Pāninian justification of students' wrong enunciations, the language did not evolve. But the poets took advantage of the recursive opportunities of the grammar to coin syntactic innovations, like arbitrarily long nominal compounds.

Thus Sanskrit is to a certain extent an artefact of linguists (Pāṇini and his predecessors and successors) who fixed a language regularizing high-register Prakrit usage, with a supplement to accommodate Vedic productions that had become archaic in usage. Since the grammar defined the sense as well as the phonetic form of meaningful enunciations (the rules indeed manipulate signs in the sense of de Saussure [13]), the language could actually be used as a knowledge-representation language. Sanskrit emerged as a formidable tool for intellectual debate, shaping sophisticated tools for dialectical argumentation, such as the *Navyanyāya* relational language.

This is the reason why Sanskrit acquired an enormous prestige for learned

debate, and even Buddhist scholars, who had initially chosen Pāli for recording their canonical texts because it was close to Magadhī, had switched to Sanskrit for their teachings by the time of Nāgārjuna (2nd century).

Thus Sanskrit was essentially fixed by Pāṇini, and further grammatical tradition was really deepening the understanding of the language but not deviating from its generative framework, like the work of Bhartṛhari (7th century), and the setting of the final complete commentary *Kāśikāvṛtti* [25] of the *Aṣṭādhyāyī*. Later grammars such as the *Siddhāntakaumudī* are merely simplified versions of the *Aṣṭādhyāyī* with a different grouping of rules to facilitate learning by beginners.

In this framework, the notion of dictionary must be revisited. After all, since Pāṇini gives structural rules for defining any nominal stem from a stock of verbal roots, it looks like a lexicon giving the meaning of verbal roots (of the order of 1000) would suffice. Actually, the tradition developed mostly specialized dictionaries, listing synonyms, homonyms, and specialized vocabularies. The main Sanskrit dictionary used in the tradition is the *Amarakoṣa* of Amarasiṃha, a Buddhist scholar at the court of king Vikramāditya (Candragupta II, 4th century). In 1000 stances, the dictionary defines 9000 word types, organized as a thesaurus expressing an ontological structure of notions. This dictionary, which gave rise to dozens of commentaries, is still in use by traditional scholars.

The rich lexicographic tradition of ancient India is detailed in [20, 15, 24].

2 Modern Sanskrit dictionaries

In analysis from written corpus, the reader is faced with a phonemic enunciation where word boundaries are lost in *sandhi*, the final phonetic smoothing that is applied last in the grammar operations to obtain a continuous verbal enunciation. Even if this deciphering is effected (in the so-called *padapāṭha* word list form), both inflexion and structural morphology rules must also be undone, a highly non-deterministic task. Thus foreigners interested in Sanskrit, but who did not have the proper traditional teaching, had to compile dictionaries of nominal and verbal stems. This led in the 19th century to a flurry of Western Sanskrit dictionaries, together with grammars generally organized on the model of Latin grammars.

Horace Wilson, a British citizen working at the East India Company mint in Calcutta, compiled the first bilingual Sanskrit-English dictionary in 1819 from various Indian lexicons; it was extended into its second edition in 1832. This work was soon superseded by the colossal *Sanskritwörterbuch* of Otto von Böhtlingk and Rudolf Roth in seven volumes, published at St Petersburg between 1853 and 1876. Böhtlingk was the editor and translator of the *Aṣṭādhyāyī*. He also produced a modified version of the dictionary in 1879 to 1889. Carl Cappeler used this material to produce a one-volume Sanskrit-German dictionary in 1887, followed by its translation as a Sanskrit-English dictionary in 1891 [3].

Monier-Williams, who succeeded Wilson at the Boden Chair of Sanskrit at Oxford University in 1860, compiled from the St Petersburg dictionaries a one-

volume Sanskrit-English dictionary, published in 1872. A later revised edition was published in 1899 with collaboration by Ernst Leumann and Carl Cappeller [18]. This 170,000 entries dictionary is to this date the main reference work on Sanskrit lexicography in the English language.

Vaman Shivram Apte published in English a number of Sanskrit dictionaries, and notably the Practical Sanskrit-English dictionary (1890, revised and enlarged in 1912), still printed today [1]. This is at this date the most popular Sanskrit-English dictionary in India. It is to a certain extent superior to Monier-Williams' in its documenting Pāninian etymologies, but it suffers from density imbalance: words staring with vowels (initial in the standard Sanskrit collation order) are denser than the ones starting with consonants.

We must add to this list a copious Sanskrit encyclopedia of Indian culture, the *Vācaspatya* by Tārānātha Tarkavācaspati, a Bengali erudite. This 5442 pages thesaurus in 6 volumes also suffers from imbalance, very complete until letter 'p', and more terse for the rest of the alphabet. Another encyclopedic work by a group of Bengali scholars headed by Raja Radha Kanta Deva is the *Śabdakalpadruma* in 5 volumes. Both works, first published at the end of the 19th century, are still reprinted.

Most of those 19th century works have been digitalized in photographic form. Many of them like Monier-Williams' are obtainable as searcheable XML databases at the Cologne University Web site[1].

A number of smaller lexicons were given as appendices of methods for learning Sanskrit ([2, 16]) or as vocabularies of selections of texts for learners (such as Lanmann's reader [17], meant as a companion to Whitney's grammar [26]). Whitney, who held the Yale Chair in Sanskrit, produced in 1885 a catalogue of roots and primary nominal forms [26] that is an indispensable tool to Western Sanskritists.

In contrast, the 20th century saw a lot less lexicographic activity around Sanskrit. In 1932, N. Stchoupak, L. Nitti and L. Renou published a *Dictionnaire Sanskrit-Français* of about 50,000 entries in 3 volumes, reprinted in 1987 as one volume of 900 pages [23]. F. Edgerton published a Buddhist Hybrid Sanskrit Grammar and Dictionary in 1953. More recently, Klaus Mylius published a 905 pages Sanskrit-German and German-Sanskrit dictionary in 2005 [19]. Oscar Pujol published in 2006 an encyclopedic Sanskrit-Catalan dictionary of 60,000 entries in 1328 pages [21].

In East Asia, Sanskrit studies were mostly motivated by the transmission of Buddhist works. But the diffusion of Sanskrit was impedimented by the lack of a phonetic system of writing. Thus Sanskrit and Pāli studies in China and Japan were contingent upon the learning of the *Siddham* system, starting with education in the syllabic writing system *Siddhamātṛkā*, evolved from the older *Brahmī* script. This writing system, at the origin of the Tibetan, Bengali and Burmese scripts, was the initial medium to transmit Sanskrit and Pāli in writing in China. In 1928 the Japanese Buddhist scholar Ogiwara Un'rai exposed a plan for a Japanese Sanskrit dictionary that was completed after his death. The full

[1] `http://www.sanskrit-lexicon.uni-koeln.de`

16 volumes work was released in 1974 as a 106,000 entries dictionary [22].

In India, a titanic lexicographic effort was undertook under the direction of S. M. Katre in 1948 at the Deccan College in Pune. About 1500 Sanskrit works were used to build a scriptorium of more than 10 million slips, recording citations for about 2 million entries. The resulting Encyclopaedic Dictionary of Sanskrit on Historical Principles had its Volume one (719 pages in Royal in-quatro) printed in 1978, under the general editorship of A. M. Ghatage. After publication of Volume 9 in 2011, it was decided to reorganize the production of smaller volumes, starting at Volume 28 (each previous volume being counted as 3 tomes). The last issued volume at the time of writing is Volume 30, published in 2014. It ends at page 5408 with entry *apramātvānadhikaraṇatva* (the state of being what is not the locus of the invalidity or the falsity of knowledge). Thus only about 6% of the vocabulary has been released over 70 years, casting doubts on its eventual completion under paper form.

Finally, let us mention that numerous more specialized lexicons - covering tantric material, architecture, *nyāya*, *āyurveda*, etc. have been produced both in India and abroad.

Around 2000, several international efforts were started in Sanskrit Computational Linguistics. This induced the design of Sanskrit lexicography databases linked with grammatical tools. The next section describes in detail one of these efforts.

In 2010 a Sanskrit Wordnet was designed at Indian Institute of Technology Bombay under the direction of Pr Malhar Kulkarni. It linked to a previous Hindi Wordnet implemented under the direction of Pr Pushpak Battacharyya, following the concepts of the English Wordnet of Princeton University. A Wordnet database is not directly a human-readable lexicon, though, but a computational tool allowing the navigation in a network of lexical items linked by relations of synonymy, antinomy, hypernomy, etc.

Several websites on the Internet propose Sanskrit digital dictionaries obtained by crowd sourcing, and thus suffering from inconsistencies and mistakes.

3 The Sanskrit Heritage digital dictionary

The author started working on a Sanskrit-French lexicon as a personal project in 1994. Initially, it was meant as a mere directory of the main notions of Ancient Indian Civilisation, as articulated in the classical Sanskrit language. But soon the complex morphological constructs of Sanskrit induced a study of the grammar of the language, and specially of its morphology. The original lexicon was written in LaTeX, with careful use of macros aligned with the internal structure of lexeme entries. By 2000 the dictionary had acquired 10,000 entries, and its PDF reached 250 pages. Its development was becoming awkward and hard to maintain, with mixture of Sanskrit, French and structure markers.

It was then decided to reverse-engineer the TeX text into a formally structured document, that could be compiled both in PDF book form, and as a hypertext HTML document. This incurred the specification of an *abstract syn-*

tax of the dictionary as a sorted algebraic structure decorated with hyperlinks. This structure proved to be robust enough in the ensueing continuous accretion. The elaboration of this structured lexicon is documented in [8, 9, 10]. In November 2003 the Sanskrit Heritage dictionary was released on Internet as the first hypertext Sanskrit dictionary[2].

At this point the project arose of using this dictionary as a generative lexicon for a computational linguistics platform. To this end, a software library 'Zen' was designed for general manipulation of lexicons, automata and finite-state transducers in the functional programming language Ocaml[3], well suited for algebraic computations in a strictly typed setting. The Heritage dictionary structure was then used as a generative device, producing a lexicon of nominal stems and verbal roots, informed with their morphological parameters. General inflexion and conjugation paradigms (*vibhakti*) were then implemented, and thus the stem tables could be compiled into databases of inflected forms, available for stemming purposes. The basic paradigms of classical Sanskrit grammar were achieved by 2003, and the inflected forms databases were then released as free linguistic data in XML form. These databanks have been used worldwide by various groups as a starting point for various computational tools for Sanskrit.

In parallel, the problem of lexical analysis of continuous Sanskrit text was attacked, in view of segmenting phonemic streams into their word components. This involves the inversion of the sandhi phonetic smoothing operation used to transcribe continuous utterances in written notation. This complex problem was modeled as finite-state relational programming. This formalism proved to be successful, and a provably correct segmentation algorithm was implemented and published in 2005 [11, 12]. This relational programming methodology was further extended by Benoît Razet's PhD research in the setting of effective Eilenberg machines, an elegant algebraic formalism generalizing the finite state machines used in the theory of formal languages.

All these ingredients were linked together in the Sanskrit Engine software, released on Internet as a set of Web services interconnecting the hypertext dictionary with the grammatical tools. A Reader assistant allowed the user to display the various segmentations of a given sentence, together with their morphological taggings. A parser was designed to trim the possibly enormous set of solutions to a smaller number of solutions respecting semantical dependencies.

A collaborative effort was then started with the Sanskrit Studies Department at the University of Hyderabad and the Sanskrit Library association, leading to the conference series ISSCL (International Symposium on Sanskrit Computational Linguistics) which became the meeting place for researchers interested in this emerging new technological field. Interoperability of various tools was demonstrated, such as the use of the Heritage Engine segmenter as the lexical component of Amba Kulkarni's dependency parser, or the invocation of the Heritage Reader from the Sanskrit Library digitalised corpus, decorated with proper invocating links. This collaborative effort was presented at COLING

[2] `http://sanskrit.inria.fr`
[3] `http://ocaml.org`

2012 as a collective distributed platform for Sanskrit processing [7].

In 2011 Pawan Goyal spent a postdoctoral year in the Rocquencourt Inria lab to work on the Sanskrit project. This collaboration led to the design and implementation of a graphical interface with proper sharing of the segmentation solutions, a much needed improvement to the Reader. This interactive interface was published in 2016 [6].

In summer 2017 Idir Lankri, a Master student, realised a prototype corpus manager, allowing the development of tagged corpus grammatically informed by annotators using the Heritage Reader as a tagging tool. This work was presented at the World Sanskrit Conference in July 2018 [14]. A first use of the corpus manager is under way, marking all the citations from the Sanskrit Heritage dictionary as analysed corpus reference through hyperlinks.

At the time of writing (December 2018), the dictionary has 32,500 entries. Its PDF book form reaches 950 pages. The Sanskrit Heritage Platform, as well as its data companion the Sanskrit Heritage Resources and its library the Zen toolkit, are developed and distributed as open-source Git repositories available at the Inria Gitlab hub `https://gitlab.inria.fr/`.

4 Design principles of a modern dictionary

Sanskrit dictionaries printed as books are historical artefacts, and future dictionaries will be sophisticated hypertext databases linked to grammatical tools. Human readable dictionaries will merely be renderings of particular views of the lexical databases. Sanskrit corpus will progressively shift from mere text files (possibly XML banks consistent with TEI guidelines) into structured text representing its analysis at various levels (*padapāṭha* segmented form, morphologically tagged text, possibly semantically tagged dependency structures, etc.) Such representations will be linked to corresponding entries in the dictionary. Thus dictionary lookup will be alleviated for readers using those tools to understand the corpus, with the lexicon just one click away. Conversely, citations illustrating usage of dictionary entries will be themselves represented in this structured fashion. All these functionalities are already implemented in the Sanskrit Heritage dictionary, with its Reader graphic interface and its Corpus tagging mode.

Computer technology may also be put to use for lexicon acquisition. The traditional workflow of a lexicographer is to collect vocabulary from a reference set of texts, together with samples of their usage in the form of idiomatic expressions, colloquial use, collocations, citations, etc. When this collecting phase is over (it took 25 years to the editors of the Pune dictionary to record the more than 10 million slips of their paper scriptorium), the dictionary editing proper iterates the typesetting of these records in a long alphabetically ordered sequence (in the case of the Pune dictionary, it took 40 years to edit 6% of the material). At the end of the whole process, additions and corrections will appear in the form of errata sections hard to consult. Modern electronic editing tools have rendered obsolete this traditional workflow. Digital texts are dynamic entities,

which may be acquired and corrected asynchronously. The whole 2-step process is useless when text analysis tools exist: missing entries will be reported by the electronic readers, leading to either further lexicon acquisition or corrupted text amendment. Thus a virtuous circle is established, where both the lexicon and the corpus are mutually supporting and cross-checking each other.

In our approach, we followed a spiral development, starting from a small lexicon with plain translations of common vocabulary, easily obtained from primers. This electronic document, properly structured, was then progressively improved by additions and corrections, when corpus examination revealed incompletenesses. Since the dictionary is the generating data for our electronic reader tool, such incomplenesses could be revealed mechanically, by parsing digitalized corpus with our lexicon-driven segmenter. Thus the dictionary data and the reader assistant software help improve each other. This shows the importance of using integrated computational linguistic tools for the bootstrap of a dictionary, and its continued acquisition. Furthermore, modern versioning tools such as Git [5] allow the smooth development of both the linguistic data and its associated software within a well-controlled cooperative effort between scholars.

This said, there are still many ways to structure dictionary entries, and the historical Sanskrit dictionaries diverge widely on this respect. For instance, it seems natural to have some hierarchy of entries. A compound such as *dharmakīrti* could appear as a sub-entry of entry *dharma*, recording its compounding with stem *kīrti*. In this spirit, Monier-Williams' dictionary is structured with 4 levels of hierarchy. The Heritage dictionary allows an arbitrary depth of compounding/suffixing. In contrast, the Pune dictionary is totally flat, every nominal stem is listed independently, even when it is a multi-component compound. At the other extreme, some lexicons, such as Bergaigne's, enter all verbs and primary derivatives derived from a root by listing preverbs into the root entry. This makes it a difficulty for beginners, who must learn how to recognize preverbs, and undo the sandhi to get to the root entry. Thus for participle form *praṇītam* one must guess that *pra* is a preverb, and that the root is actually *nī*, since the retroflex *ṇ* is just a consequence of (internal) sandhi. Then one may consult the root entry *nī* under its sub-entry for preverb *pra* to finally find the participial stem *praṇīta* whose accusative form is *praṇītam*. Note that all these difficulties vanish when an analysis tool such as the Heritage Reader is presented with form *praṇītam*, and immediately returns its morphological tag as

{acc. sg. n. | nom. sg. n. | acc. sg. m.}[pra-nīta {pp.}[pra-nī]]

with direct access to root *nī* in the lexicon, where *pra* is listed as a legitimate preverb for that root, presented as a direct link to verbal entry *praṇī*.

Many subtle choices may be made concerning the granularity of entries. For instance, homonyms ought to be distinguished when they are produced from distinct roots. But a nominal stem (*prātipādika*) such as *prabodhana* ought to cover both its agentive meaning (as an adjective in all genders "who wakes up") and its action meaning (as a substantive in neuter gender "awakening") in two different sections. In other words, we consider the two constructions as

polysemic variants. Similarly, a compound such as *pītāmbara* would have two sections, one for its determinative usage (*tatpuruṣa*: yellow garment, a neuter substantive) and one for its possessive exocentric usage (*bahuvrīhi*: who has a yellow garment, adjective in all genders, reifiable as a masculine substantive to designate Viṣṇu). However, a stem such as *nirvācya* ought to be recorded as two distinct entries, one analysed as the non-compositional *prādi* compound *nir-vācya* (what should not be discussed), another one as the (compositional) future passive participle of the causative conjugation of verb *nirvac* (what ought to be explained), even though they both derive from root *vac*. Monier-Williams similarly distinguishes the two formations, but the two entries wind up in distant places of the dictionary, making it hard to choose between them.

In any case, the granularity ought to be correlated to the Pāninian theory. Formations are determined by suffixes which determine not only the signifier (the phonemic string) but also the signified (its meaning, such as agent of its etymological root). Subtle semantic distinctions in the grammar (like whether the agent is occasional or habitual/professional), will help in organizing the polysemic structure, but should not give rise to separate morphological sections. The Heritage dictionary is getting progressiveley aligned with a Pāninian-consistent account of morphology. This requires a finer abstract structure of lexical entries. This also induces extra work in the morphological justifications - every nominal stem must be justified either by the sequence of Pāninian *sūtra*s necessary for its production (*prakriyā*), or by an entry in the *uṇādisūtrakośa*.

One important consideration is whether one intends to use the lexicon as a mere dictionary of common words, giving definitions by proper paraphrases in the target language, meant as a tool for learners of Sanskrit interested in the language structure but not in its cultural context, or whether on the contrary one wants to express the usage of words in the Sanskrit Indian culture, with all their metaphoric and semiotic richness evoked in literature. Simple dictionaries for beginners are often of the first kind. For instance, Bergaigne does not list *śiva* in his lexicon "in order not to burden the student with Indian matters". Whereas experts doing philological work will need to have detailed explanations of technical usage in the various traditions, and for instance distinguish technical definitions of words like *dharma* according to the various philosophical schools. In that case the dictionary will be more like an encyclopedia of Indian culture.

One point in case is proper names. Sanskrit proper names are in general derivable as common nouns. The written language, whether in the traditional *devanāgarī* or in other Indian scripts, does not indicate proper names by typography, like the capitals of Roman scripts. Thus it us unavoidable to mix proper names information with their Sanskrit literal originals. Thus information about *Śiva* will be found as an encyclopedic subentry of the common adjective *śiva* "favorable". How much of such information should be provided is a hard question. For instance, Monier-Williams lists thousands of titles of literary works as entries bearing the bare indication "N. of wk." without mention of author, origin or datation, which is rather frustrating. In contrast, the Heritage Dictionary lists only works given with their authors and dates, in as much as this data is available as consensual opinion of competent scholars. The hypertext structure

then helps in navigating through the proper information.

This explains the large size of the Sanskrit Heritage database, compared to its rather modest number of entries. It is really a small encyclopedia on ancient Indian culture through the mediation of Sanskrit. Organisation of this material is much more subjective than the literal meaning of common words. Consequently, crowd sourcing of such information is ill-advised, since uniformity and consistency are of paramount importance in these matters. Better have a biased view of the concepts than an inconsistent one.

Encyclopedic entries raise many issues about classification (should it correspond to the concepts of the target language ontology, or rather to the ones of the Indian tradition?), location and datation, authenticity of source information, etc. Natural sciences concepts, such as plants and animals names, should both relate to modern standards of biological nomenclature, and at the same time inform about relevant semiotics in poetry, as well as traditional usage in pharmacopeia. We shall not discuss further these side-issues.

Many more issues arise in the making of a lexicographic database, such as the structure of meanings, their ordering (by frequency, by etymologic/metaphoric shifting, by diachonic development, by technical distinction according to philosophical schools), etc. Indications of usage, from idiomatic expressions to citations of full sentences, are also a matter of design. In the Heritage dictionary, such usage indications are appendices to a stem entry, in several flavors (idiomatic uses of forms of the entry, phrases documenting frequent collocations, finally corpus citations). In more complete dictionaries such as Apte's, such usage indications are listed along the individual senses, as is more appropriate. In digital dictionaries, information may be structured in more sophisticated ways than paper dictionaries. Most of the information may be hidden, and available by progressive clicking, thus not cluttering the contextual information display if the user does not need it. Thus learner modes could allow short display of the most frequent senses, uncluttered by encyclopedic information.

Conclusion

We discussed a number of issues relevant to the design of a Sanskrit dictionary. There is an important tradition of Sanskrit lexicography, first by the ancient Sanskrit grammarians, and in the last two centuries as outcome of Western indological research. Future Sanskrit lexical databases will profit of this heritage, but should provide more sophisticated usage for their users through digital technologies.

References

[1] V. S. Apte. *The Practical Sanskrit-English Dictionary: Containing Appendices on Sanskrit Prosody and Important Literary and Geographical Names*

of Ancient India. 3rd edition, 1965, Motilal Banarsidass, Delhi, 1890, revised and enlarged 1912.

[2] A. Bergaigne. *Manuel pour étudier la langue sanscrite*. F. Vieweg, Paris, 1884.

[3] C. Cappeller. *A Sanskrit-English Dictionary, based upon the St. Petersburg lexicons*. Trübner, Strassburg, 1891.

[4] G. Cardona. *Pāṇini, A Survey of Research*. Mouton, The Hague, 1976.

[5] S. Chacon and B. Straub. *Pro Git*. Apress (available as `https://git-scm.com/book/en/v2`), 2014.

[6] P. Goyal and G. Huet. Design and analysis of a lean interface for Sanskrit corpus annotation. *Journal of Linguistic Modeling*, 4(2):117–126, 2016.

[7] P. Goyal, G. Huet, A. Kulkarni, P. Scharf, and R. Bunker. A distributed platform for Sanskrit processing. In *COLING*, pages 1011–1028, 2012.

[8] G. Huet. Structure of a Sanskrit dictionary. Available as `pauillac.inria.fr/~huet/PUBLIC/Dicostruct.ps`, 2000.

[9] G. Huet. From an informal textual lexicon to a well-structured lexical database: An experiment in data reverse engineering. In *Working Conference on Reverse Engineering (WCRE'2001)*, pages 127–135. IEEE, 2001.

[10] G. Huet. Design of a lexical database for Sanskrit. In *Workshop on Enhancing and Using Electronic Dictionaries, COLING 2004*. International Conference on Computational Linguistics, 2004.

[11] G. Huet. A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *J. Functional Programming*, 15,4:573–614, 2005.

[12] G. Huet. *Themes and Tasks in Old and Middle Indo-Aryan Linguistics, Eds. Bertil Tikkanen and Heinrich Hettrich*, chapter Lexicon-directed Segmentation and Tagging of Sanskrit, pages 307–325. Motilal Banarsidass, Delhi, 2006.

[13] G. Huet. Sanskrit signs and Pāṇinian scripts. In A. Kulkarni, editor, *Sanskrit and Computational Linguistics*. D.K. Printworld, New Delhi, 2016.

[14] G. Huet and I. Lankri. Preliminary design of a Sanskrit corpus manager. In G. Huet and A. Kulkarni, editors, *Computational Sanskrit & the Digital Humanities*. D.K. Printworld, New Delhi, 2018.

[15] V. N. Jha, editor. *Proceedings of the National Seminar on the Art of Dictionary Making in Ancient India*. Center of Advanced Study in Sanskrit, University of Pune, 1997.

[16] F. Knauer. *Ucebnik' sanskriskago yazyka*. W. Dragulin, Leipzig, 1908.

[17] C. R. Lanmann. *A Sanskrit Reader*. Boston, 1906.

[18] M. Monier-Williams, E. Leumann, and C. Cappeller. *A Sanskrit-English Dictionary: Etymological And Philologically Arranged With Special Reference To Cognate Indo-European Languages*. Asian Educational Services, 1999.

[19] K. Mylius. *Sanskrit-Deutsch / Deutsch-Sanskrit Wörterbuch*. Harrassowitz Verlag, Wiesbaden, 2005.

[20] M. M. Patkar. *History of Sanskrit Lexicography*. Munshiram Manoharlal, New Delhi, 1981.

[21] O. Pujol. *Diccionari Sànscrit-Català*. Enciclopedia Catalana, Barcelona, 2006.

[22] M. Sato. Sanskrit dictionaries in Japan: Focusing on the works of Ogiwara Un'rai. In L. Wei, editor, *Research on the Language and Script in Buddhist Sutras*. Hangzhou Buddhist Academy, 2019.

[23] N. Stchoupak, L. Nitti, and L. Renou. *Dictionnaire Sanskrit-Français*. Librairie d'Amérique et d'Orient, 1987.

[24] C. Vogel. *Indian Lexicography - Revised edition*. P. Kirchheim Verlag, 2014.

[25] Vāmana and Jayāditya. *Kāśikā (A commentary on Pāṇini's Astādhyāyī)*. Sanskrit Academy, Osmania University Hyderabad, 2008.

[26] W. D. Whitney. *Roots, Verb-forms and Primary Derivatives of the Sanskrit Language*. Motilal Banarsidass, Delhi, 1997. (1st edition 1885).