Smart lexicography for low-resource languages: lessons learned from Buddhist Sanskrit and Classical Tibetan.

Ligeia Lugli

SOAS University of London, Thornhaugh Street, London WC1H 0XG, room 339
E-mail: ll34@soas.ac.uk

Abstract

Traditional lexicography requires titanic efforts and enormous resources. For many languages, such resources have never been available. As a result, they received only limited lexicographic coverage. Today, these languages can take advantage of many of the same digital tools and strategies that have simplified and expedited dictionary-making for mainstream languages. Yet, the resource gap remains evident even in the digital era, with basic corpus processing tasks that lie at the foundation of contemporary 'smart lexicography' still constituting a challenge for many under-resourced languages.
Drawing on my own experience in Sanskrit and Tibetan lexicography, this paper aims to offer some guidance as to the advantages and limitations of the application of smart lexicography to under-resourced languages. In particular, this paper suggests that in order to optimise resources, it may be advisable to prioritize high-quality lexical annotation of the corpus over highly curated dictionary entries, and to let digital tools take care of the lexicographic representation of the annotated linguistic information.
Keywords: automated lexicography, GDEX, Buddhist Hybrid Sanskrit, Tibetan.

1. Introduction

This paper serves two purposes. On the one hand it provides a progress report of two ongoing lexicographic projects, (1) a Buddhist Sanskrit lexical resource called The Buddhist Translators Workbench commissioned by the Mangalam Research Center (Berkeley, CA), and (2) a diachronic valency lexicon of Tibetan verbs, which is being developed at SOAS (University of London) within the AHRC-funded project Lexicography in Motion. On the other hand, this paper outlines strategies for applying smart lexicography to low-resource languages.

Smart lexicography is intended here as an optimally efficient cooperation between human lexicographers and machines, whereby all task that can be automated are delegated to computers, while lexicographers focus on points of curation that require human judgement. This includes re-using pre-existent dictionary content and ensuring that any new human-curated output can in turn be re-used by other project or in subsequent iterations within the same project.

What constitutes a 'low-resource language' is more difficult to define. Low is a fundamentally relative concept, as it acquires meaning only relative to its antonym 'high'. Languages can be considered low-resource only when compared with high-resource

languages, like English or other major spoken languages that tend attract much study, funding and technological development. In this paper, I use the expression 'low-resource languages' to indicate those languages for which computational and human resources are insufficient to take full advantage of state of the art of automated or semi-automated lexicographic workflows.

Many different reasons may limit the ability to apply automation to lexicographic tasks. For the projects discussed here, one crucial obstacle has been the difficulty of producing suitably annotated corpora quickly. Sadly, Rundell and Kilgarriff's assertion that "the timescale for creating a large lexicographic corpus has been reduced from years to weeks, and for a small corpus in a specialised domain, from months to minutes (Rundell and Kilgarriff 2011)" does not apply to the languages considered here. The main problem for these languages has been generating sufficient manually annotated data to develop reliable NLP pipelines for corpus pre-processing. Few people have the adequate skills to create the amount of annotated data necessary to train Machine Learning-based models, or even to test rule-based systems. Moreover, these people are usually highly skilled, not easily amenable to the dull routine of corpus annotation and required for more sophisticated lexicographic tasks.[1]

Fortunately, the unavailability of large amounts of training data needs not preclude the application automation to the lexicography of low-resource languages entirely. It does however impose significant limitations on the scope of such application and the results that can be achieved through it.

A key to the adoption of smart lexicography for low-resource languages lies in the re-conceptualisation of the dictionary product and of its core design principles. Good lexicographic practice dictates that entries are designed primarily to meet the needs of the dictionary prospective audience, or 'market' (Atkins and Rundell 2008, Ch. 2; Landau 2001, 343). While this is undoubtedly a commendable approach, when working with low resource languages much is to be gained if the needs of the lexicographic team take primacy over those of the audience. As this paper will show, ambitious microstructures designed to fulfil the audience needs may slow down the progress of small teams working on low-resource languages to unsustainable levels. By contrast, investing the lexicographers' linguistic expertise to create annotated data future can lead to faster and more rewarding results. This is because annotated data is inherently versatile. It can be immediately displayed to users in the form of a lexical database or minimally curated 'proto-dictionary', it serves to develop NLP pipelines and can be later re-used to create

[1] This is critical issue for historical languages like Buddhist Sanskrit and Classical Tibetan, for which no active speakers are available. Contemporary low-resource languages may pose different challenges; cf. Nasiruddin 2013 who sees Machine Learning as promising for under-resourced languages for which crowd-sourcing solutions are available.

full-fledged dictionaries (cf. Pajsz 2009, Atkins and Rundell 2011 and Mianáin and Convery 2014). This strategy fits the definition of smart lexicography given above insofar as it constitutes an optimally efficient cooperation between lexicographers and computers given the available human and digital resources.

This is the general strategy we have adopted, to varying degree and with different practical solutions, in the two projects discussed in this paper.

2. The Buddhist Translators Workbench

2.1 Project overview

The project was commissioned by the Mangalam Research Center in 2012, with an eye to provide translators with useful lexical information about key Sanskrit Buddhist vocabulary. The primary aim of the project was to help translators achieve a nuanced understanding of selected Buddhist vocabulary and, ideally, move away from the overly terminological renditions and calques that often characterise English translations of Buddhist Sanskrit Texts (Griffiths 1981). Two features were deemed essentials to achieve this goal.

First, the dictionary would have to be corpus-driven. Semantic descriptions and lexico-semantic relations should be derived from the corpus rather than from traditional interpretation. This decision was at odds with the perceived needs of a sizeable portion of our intended audience, who was primarily interested in historical normative lexicography and asked that we derive our content from traditional Buddhist definitions found in ancient treatises and present it in the form closer to encyclopaedic articles than dictionary entries (Lugli 2019). Dauntingly, introducing corpus lexicography in the field of Buddhist Sanskrit also required building a suitable corpus from scratch. Buddhist Sanskrit is a non-classical variety of Sanskrit, sometimes referred to as 'Buddhist Hybrid Sanskrit' (Edgerton 1953), that is especially difficult to segment and has hardly received any attention from the NLP community until very recently.[2] With no computational tools available to process Buddhist Sanskrit, we opted for working with a very small unprocessed corpus consisting of 33 Buddhist Sanskrit texts dating from the first half of the first millennium CE and belonging to various traditions and text-types. The choice of the texts was largely determined by the quality of the available digital editions and the availability of translations. Given the amount of manual labour involved in retrieving and analysing corpus examples for each lemma, starting on such a small corpus seemed a justifiable choice.

[2] See Lugli 2018 and forthcoming, as well as Handy 2019.

Second, detailed lexical analysis would be presented in narrative form together with sense-descriptions, examples and short etymological overview. As a compromise between our intended mission and our audience's requests, we decided to open our entry with a rather lengthy narrative description of the headword that would explain the relationship between its general and specialised uses in a format akin to a miniature essay. Great efforts were invested in the design and implementation of a granular microstructure that would provide users with the information necessary to gauge the semantic versatility of key Buddhist words in context, and appreciate their relationship with semantically and etymologically related words. Since our intended audience comprised both seasoned scholars and students we also took care of presenting the information in a way that would satisfy both user groups. The entry would provide our analysis of a lemma while at the same time also offering users the opportunity to conduct their own analysis based on an extensive range of examples extracted from the corpus. All the examples found in the corpus would be semantically categorised, but only those judged to be most illustrative of a sense or construction would be rendered in English.[3] For each sense of a lemma, the entry would also provide a 'contrastive section' with examples illustrating the relationship between the lemma and semantically or etymologically related words in context.[4]

2.1.1 Problems

Several entries were produced using the microstructure outlined above. Work was progressing extremely slowly and it gradually became clear that the amount of labour required to prepare an entry was simply not sustainable. This was partly due to the large amount of curated information that each entry required. The translation of all the relevant examples alone typically took several days. Yet, what proved to be really unsustainable was the kind of workflow that the essay-like entry required–and its tolerance for lack of systematicity. Combined with the training background of our lexicographic team, this workflow led to catastrophic results.

People proficient in Buddhist Sanskrit tend to have a solid philological and philosophical training, but no training in lexicography and corpus linguistics. This affects their lexicographic output in several ways. First, they are not used at looking for patterns in data and find it difficult to abstract word senses from individual citations, or spot correlations between meaning and co-text. Second, they tend to focus on philosophically interesting examples where the lemma is used in a less than typical way.[5] Third, and most important, they are used to a scholarly workflow that starts with taking notes and

[3] On the system of semantic categorisation used in the project see Lugli 2015.
[4] For more information of the principles informing the entry design, see Gomez and Lugli 2015.
[5] Cf Atkins and Rundell 2008, 52.

progresses by gradually refining these notes into a publishable piece of writing. This workflow was initially encouraged as it was thought suitable to produce the verbose entries that the project required. This proved to be the single most problematic aspect of the early phases of the project. The unstructured workflow made it difficult to monitor progress, reproduce the lexical analysis that informs an entry, or hand over an unfinished entry to colleagues whenever a contributor left. Most importantly unstructured note-taking was in no way re-usable and could not contribute to advancing the NLP infrastructure that we needed to build a lemmatised corpus.

After years of painfully slow progress, a costly lesson was learned: before staring lexicographic work (especially on a low resource language), it is advisable create a highly structured digital workflow designed to optimise resources. In our case, a good way to optimise resources was to ensure the re-usability of the lexicographers' output for both dictionary content and corpus creation.

To move from this realisation to its implementation was not easy. The idea of adopting a rigid workflow met with significant resistance and was at first rejected on the grounds that it would be too mechanical a job for postdoctoral scholars, and junior students would not have sufficient proficiency in the language to perform it accurately. Both objections are valid. It proved difficult to find collaborators who are both capable and willing to annotate Buddhist Sanskrit using a systematic workflow. Still, the time invested in searching for these people and developing a computer-assisted workflow proved a good investment.

2.2 Towards smarter lexicography for Buddhist Sanskrit

In 2017 we developed a web-based annotation tool that requires lexicographers to record syntactic and semantic information for each citation (i.e. KWIC) they analyse.[6] The corpus is still unprocessed, so the annotation tool requires lexicographers to manually segment and lemmatise the examples, mark all syntactic dependencies involving the lemma, semantically tag the lemma and its dependencies, and annotate conceptual relations between the lemma and other co-text items (e.g. cases where the concept expressed by a lemma is said to be caused by a concept expressed by another word in the sentence). Given the interpretive difficulties of the sources, lexicographers are also asked to record any uncertainty in the annotation using a four-fold typology that allows to distinguish between philological problems, textual ambiguity, disputed interpretation and personal uncertainty (Lugli 2015). Finally, the annotation process involves aligning the Sanskrit examples with their published English translation.

[6] https://btw.mangalamresearch.org/en-us/meaning-mapper/

Such detailed annotations are time consuming. Still switching to an annotation-based workflow has sped up lexicographic work of an order of magnitude compared to the unsystematic workflow we initially had. It has improved the efficiency of our in-house lexicographic training phase, facilitating our contributors to transition from a 'humanities mindset' to the adoption of corpus-linguistics methods. It has also made lexicographers' analyses more transparent and easy to check, thus drastically reducing the time allocated to revisions. Most importantly, the new workflow has enabled us to adopt an iterative lexicographic cycle whereby proto-dictionary entries automatically derived from the annotations can be made accessible to our audience before fully curated entries become available.

2.2.1 A Visual Dictionary of Buddhist Sanskrit

With the new workflow, the immediate output of our lexicographers' work on a headword is not a dictionary entry; it is a dataset containing annotated citations for that headword. This dataset can be exported from the annotation tool to several formats, including vertical, xml or CSV. Each format has its own uses. Here I will focus on the CSV format, which offers the advantage of easily lending itself to analysis through widely used statistical computing platforms such as R.

The CSV files exported from our annotation tool have one row per citation and one column per annotation field. For example, there are columns containing semantic descriptors of the headword at various levels of granularity (e.g. semantic field, sense and subsense). There are also columns for grammatical details such as gender and number, as well as several columns devoted to syntactic information. The representation of syntactic dependencies over CSV columns is somewhat clumsy, especially if compared to CONLL formats, but it is nonetheless effective. Each type of syntactic relation corresponds to a variable (e.g. 'modifies', or 'isSubjectOf') that takes as values the lemma forms of the words linked to the headword through the specified syntactic relation. The same applies to conceptual relations. The resulting CSV features 170 columns and is best explored through data visualisations.

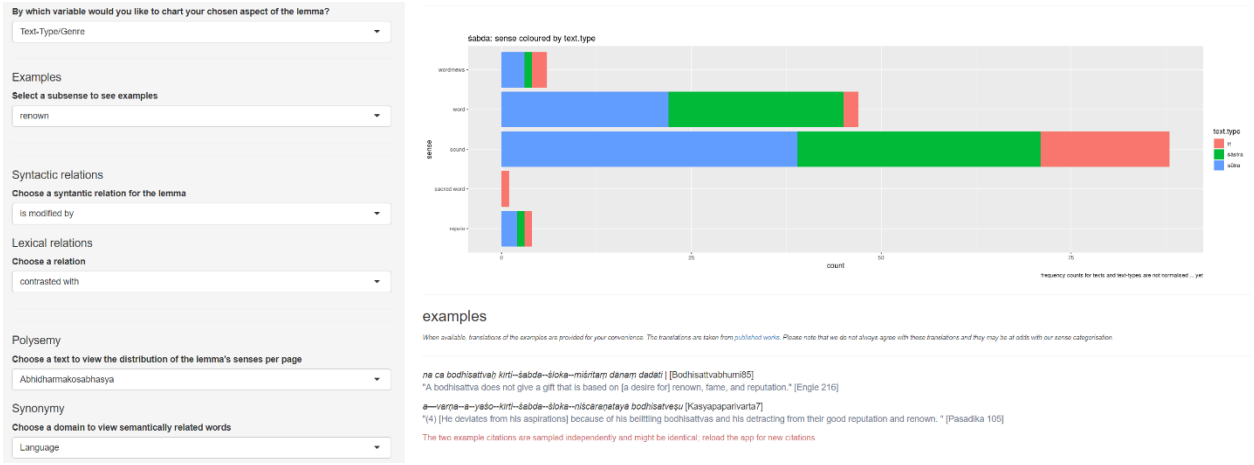
These visualisations, which we currently generate using the popular R package ggplot2, are used internally to check the consistency of the annotations. They also serve to refine the lexicographers' interpretation of a lemma in context, highlighting collocational trends and co-textual patterns that might have been overlooked while reading through the citations.

Once the dataset for a headword has been checked and the team agrees that the annotations it contains are reliable, it is merged with the datasets already created for other words and the information it contains can immediately be made available to the public via

those very same data-visualisations we used internally to refine the annotations. To this end we currently use Shiny, an R package that allows users to create web-based interactive apps with minimal programming skills (Chang et al 2018). Shiny is extremely versatile and supports data-visualisations as well as text sections, thus allowing the display of traditional dictionary content, such as definitions and examples, as well as charts.

At present, our Shiny app is a rapidly evolving working prototype called (over-ambitiously) A Visual Dictionary of Buddhist Sanskrit.[7] It opens with a shallow description of the senses and semantic domains covered by the lemma, which is automatically derived from the annotated dataset, followed by a series of data-visualizations that allow users to explore various aspects of the lemma. The top visualisation can be configured to chart most of the information contained in the annotated in the dataset, including the distribution of a headword's senses, subsenses and semantic prosody across different genres, period, traditions and periods.

Below this graph, the app displays two corpus examples where headword expresses the sense or subsense chosen by the user. The examples are accompanied by bibliographic references and, whenever possible, they are followed by a translation taken from a published translation of the relevant text. Currently the examples are randomly selected from among all the examples available for a word-sense combination. A GDEX-based system may be devised once we have a segmented and lemmatised corpus.



After the examples, the user is presented with a series of word clouds, illustrating the relative frequencies of various co-textual items that occur in the user-specified relation with the lemma. Further down, the user can visualise the distribution of word-senses in a specified text. This visualisation addresses one of the primary concerns of the original Buddhist Translators Workbench project, that is helping translators gauge the degree of

[7] https://ligeialugli.shinyapps.io/VisualDictionaryOfBuddhistSanskrit

specialization that a lemma might have in a given text and appreciate the semantic continuity that often exist between the artificially created word senses. This feature is especially useful for students of Buddhist philosophy, as it helps identify cases where the inherent vagueness of a word was exploited for hermeneutical reasons. Typically, the chart would highlight these cases by showing the deployment of different senses of the same word in close proximity. At the time of writing, the unit used to measure proximity is the page of the Sanskrit edition of the text. This is unhelpful, as the length of pages changes from text to text and thus impairs comparison of a lemma's semantic distribution across different sources, which is a desirable feature.[8] We are in the process of switching to a sentence-based measure to enable such comparison.[9]

The last visualisation that our app currently offers is a chart that categorises lemmata by semantic domain to identify near-synonyms. We will probably soon switch to a different modality of visualisation for this chart, and as soon as we will have sufficient data we intend to move away from relying on semantic annotation for this feature and we will seek to use corpus data and collocational information to detect potential synonyms.

It is important to emphasise that this app is a work in progress and has not been developed by our professional engineer. It is a conceived as a nimble tool to communicate our results to our audience in real time without incurring into additional software-development cost.

2.2.2 Future developments

We are currently creating datasets for headwords pertaining to the semantic fields of language and mental activity, with an emphasis on lemmata that cover both semantic fields. Once we complete datasets for all the words in these semantic fields, we will start a new iteration of the lexicographic process and craft human-curated descriptions of the words to replace the shallow automatically generated summaries that currently open the entries. Once the curated descriptions are in place, our lexicographic team will move on to annotating citations for words related to a new semantic field, while contributors with no specialised knowledge of Buddhist Sanskrit will be tasked with filling in our original work-intensive microstructure with the data annotated by the lexicographers. This allows the 'real' dictionary to keep growing at reduced cost. Once the datasets for one semantic field are deemed complete, they will also be made available to the public in CSV, CONLL and xml format for re-use in other projects.

This iterative model allows us to concentrate our very limited human resources on one

[8] I am grateful to Ammon Shea for suggesting this feature.

[9] This is not without problems, as 'sentence' is not a straightforward concept in our sources, and some differences in the division of text into sentences may occur from text to text.

task at the time, first annotation and then lexicographic curation, while simultaneously enabling our audience's access to lexical analysis at an early stage. It also allows us to work towards the development of a fully processed corpus. The manually segmented citations have been used to develop a rule-based segmenter and lemmatiser that is currently being used to automatically process our corpus (Lugli forthcoming). The manually annotated dependencies are also being used to test a Sanskrit sketch grammar for use in the Sketch Engine that has been developed by the present author. This sketch grammar is designed to infer syntactic relations from a segmented corpus, without need of PoS tagging or dependency annotation. As it relies on morphology only, it cannot achieve the same level of delicacy as the manually annotated citations. Yet, the ability to infer even the most basic syntactic relations (e.g. verb's subject and object) automatically would constitute a significant advance for Buddhist Sanskrit corpus linguistics. If the automatically inferred syntactic relations will prove sufficiently accurate, we shall be able to further streamline our lexicographic work by limiting annotation to semantic information. In the future, semantic tagging could also be automated, but this avenue has not been explored yet within the project.

3. Lexicography in Motion: a Tibetan verb valency lexicon

The context of the diachronic Tibetan verb lexicon project differs significantly from that of the Buddhist Translators Workbench. This project builds on extensive previous work on Tibetan NLP. It disposes of at least two PoS taggers and lemmatizers (Garrett, Hill and Zadoks 2014, and Meelen and Hill 2017), as well as of a large tokenised, lemmatised and PoS corpus (Meelen, Hill and Handy 2017). It also benefits from pre-existent high-quality dictionaries, including works devoted entirely to Tibetan verbs (Hackett, 2019, Hill 2010). Moreover, the team possesses expertise not only in the Tibetan language, but also in professional lexicography and computational linguistics. Yet, this project, too, partakes of some key difficulties characteristic of lexicography of low-resource languages, especially for older diachronic strata–which are the focus of the present discussion. Even though pre-processed corpora for these strata of the language exist, they do not possess the layer of annotation required for our lexicographic purposes. The main research goal of the project is to shed light on verb argumentation patterns through corpus evidence. To this end, the lexicon relies on a annotation system for syntactic dependencies that distinguish between twelve types of arguments.[10] Few researchers in our team possess the necessary level of language proficiency to carry out the dependency annotations or check the output of automatic parsers. They are the same people who were initially tasked with creating the dictionary content.

This creates intra-project competition for human resources, as the same team-members are needed for NLP and corpus development on the one hand, and for lexicographic

[10] For details, see https://tibetan-nlp.github.io/lim-annodoc/deprels.

curation on the other. We planned to address this problem by tasking these researchers with corpus annotation first, and with lexicographic editing later on. The idea was that once a critical mass of manually annotated data was achieved, dependency annotation could be automated. In the meantime, the rest of the team would prepare the microstructure of the dictionary and ready a dictionary writing schema for the lexicographers to use as soon as the corpus was ready.

The theory behind this plan is sound. In practice, reaching a critical mass of manually annotated sentences and developing a reliable automated dependency annotation has been taking most of the team's time and energies, leaving very little room for lexicographic curation of dictionary entries. As a result, the automation of lexicographic tasks has acquired a more prominent and pervasive role in the project than we had initially envisioned.

A challenge in this project is that our corpus' design is still in flux. The corpus is being built while we devise strategies for automatically extracting and displaying lexicographic information from it. Any trials and tests need to be run on the exiguous manually annotated corpus that we currently have, which amount to around one hundred thousand words. Yet, the solutions we come up with through the trials need to be scalable to the full corpus once we have it. The size of our final corpus is not set, but will ideally include several hundred million words.[11] Size is not the only difference between the corpus we are using for trials and the one on which we intend to base our final lexicographic product. The final corpus will comprise three diachronic layers, while so far we have been working only on Classical Tibetan. The dependencies annotation will be enriched with morpho-syntactic information that are currently not available, and portions of the corpus will be aligned to English translation. In brief, our strategies for automating the project's lexicographic output need to be adaptable to changes in the corpus.

3.1 Lexicographic automation for a diachronic Tibetan verb valency lexicon

In collaboration with the Sketch Engine, we have generated a sample dictionary draft from our small manually-annotated corpus of Classical Tibetan. It contains 774 entries, based on a headword list derived from existent Tibetan dictionaries. We also derived a headword list from the corpus, but this proved unsatisfactory, as it erroneously included nominalised verbal forms, due to PoS-tag ambiguity. When our full corpus will be ready, we will derive a new headword list from it and compare it with the list extracted from dictionaries to ensure that verbs not recorded in existent dictionaries but attested in the corpus will be

[11] Ideally it would comprise a 300 million word corpus of Tibetan that has been PoS tagged in recent years (Meelen, Hill and Handy 2017), plus an additional corpus of contemporary Tibetan and a small corpus of Old Tibetan that we are creating from scratch within the project.

included in our lexicon.

Our small test corpus is associated with a sketch grammar that allows verbs' word sketches to be arranged by argument structure in Sketch Engine. The word-sketch information is mapped onto our DWS entry template, which is arranged by argument structure. As DWS we are using Lexonomy, a free dictionary writing software closely connected with the Sketch Engine. Lexonomy allows users easily to edit entry templates that can be auto-populated with information from a corpus hosted on Sketch Engine (Měchura 2017). Lexonomy's out-of-the-box configuration allows lexicographers to pull dictionary examples from a Sketch Engine corpus from individual example slots in each entry. This practice requires lexicographers to manually select and add the examples to the entries, which is time consuming. To push all the examples from the corpus directly to the relevant slots in the entries seems more efficient; so we opted for this solution. This required the assistance of the Sketch Engine team and the payment of a (very reasonable) fee.

3.1.1 GDEX development for Classical Tibetan

In the dictionary draft, all examples are accompanied by full bibliographic and period metadata and are sorted using a GDEX formula that models an ideal good dictionary example (Kilgarriff et al. 2008). The main parameters of our GDEX are sentence length, absence of additional arguments beside the argument pattern to be illustrated by the example, and a reduced presence of pronouns, to avoid anaphoric references that may be difficult to interpret out of context. To filter out sentences that might be difficult to read, examples with many verbs are penalised, and so are those displaying lengthy strings of adjectives, determiners and adverbs.

Our GDEX formula was first intuitively developed on the basis of an ideal model of 'good Tibetan example sentence'. The output of the formula was then tested against 150 sentences manually rated by the lexicographers on a 0-2 scale, where 2 is a perfect example, 1 is an example that may need some manual editing, and 0 is bad example. 70% of the examples were rated 0, and only 8% were rated 2. Given the limited time the lexicographers could spare for rating examples, only two iterations of the formula have been possible so far. The formula that we have developed through these iterations is successful in promoting good examples to the top of the example list; but given the paucity of 2-rated examples it was impossible to fine-tune the formula to distinguish between 1- and 2-rated examples. It also needs improvement in filtering out 0-rated sentences. Currently, while all good examples are among the top-rated sentences, almost one third of the top-rated sentences are bad examples.

The identification of complete sentences is one of the most challenging aspect of

modelling good examples for Classical Tibetan. The corpus is divided into sentences according to Tibetan punctuation, but this does not follow the same principles as Western punctuation and is rarely indicative of sentence boundaries. Steps have been taken to include likely identifiers of final sentence boundaries in the GDEX formula. For instance, sentences ending with final particles are promoted, while sentences ending with case markers are penalised. Yet, more work remains to be done to identify initial sentence boundaries.

As it is often the case with GDEX, our current formula promotes simple sentences. These may well be user-friendly, but are not necessarily representative of the style employed in Classical Tibetan sources. For this reason, our entries will also contain examples sorted through an alternative GDEX formula that does not penalise multiple verbs, modifiers and determiners as much as the current one. It will be up to the user to choose which set of examples to peruse.

In an effort to promote at top of the example list the most representative sentences, we have also augmented GDEX sorting with argument-specific collocational information.[12] The highest GDEX-ranked example that features in the relevant argument slot the most frequent word for that argument slot is promoted to the top. Likewise, the top GDEX ranked example that has in the relevant argument slot the second most frequent word for that argument slot will occupy the second position in the example list, and so on. This is to ensure that at the top of the example list we will have typical sentences like 'to drive a car' and not idiosyncratic expressions like ‘to drive a gas guzzler’.

To be representative, the top examples also need to be drawn from a variety of sources. All the rest being equal, the sentences at the top of the example list will be taken from different texts.[13]

3.1.2 Future developments

To be useful, dictionary examples need not only be 'good' and representative, but also easy to peruse. In the case of ancient languages such as Old and Classical Tibetan, adding a translation of the examples would help in this regard. It is unlikely that our lexicographers will have time to craft such translations; so our attention has turned to the possibility of using published translation of the sources. While it may save us time, this option is not without its problems. Only a fraction of our final corpus has been translated. This leaves us with the uncomfortable choice of either limiting the selection of our top examples to the few texts that we can align with published English translation, thus not taking full advantage of the power of the integrated GDEX workflow we have devised, or

[12] Cf. Gantar et al 2016, 214.
[13] Cf. Cook et al. 2014 320-321.

risking to leave the top examples untranslated, thus compromising the user-friendliness of our lexical resource. A solution would be to allow users to decide whether or not to restrict the selection of the examples to those accompanied by translation. We have not yet investigated how to implement this feature within the Lexonomy infrastructure.

The most daunting challenge awaiting us is the addition of word senses to the entries. Currently, the entries are divided by argument pattern and not by sense. This allows us to auto-populate the entries purely on the basis of word-sketches, without recourse to automatic sense induction or sense discrimination. Senses feature in our Lexonomy entry schema as xml attributes of example elements, alongside bibliographic and period metadata. The original aim of this arrangement was to allow lexicographers to manually tag the top examples with sense labels while editing of the automatically generated dictionary draft. It now seems unlikely that the lexicographers will have sufficient time to sense-tag the examples, as their linguistic expertise is still needed to develop the dependency parsed corpus. We will therefore explore avenues to automate this aspect of the lexicographic work, too.

4. Conclusions: lessons learned

Automated lexicographic solutions can only be as smart as the language resources they rely on. Languages that lack suitably processed and annotated corpora are at a disadvantage. Especially so, if there is a paucity of people able to annotate those corpora and develop adequate NLP tools for them. Still, this is no excuse for reverting to entirely manual workflows. The lexicographers' work and output should be designed to serve more than one purpose, so that beside building dictionary content it also feeds in NLP research and contributes to the creation of better corpora, which will, in due course, enable faster lexicographic workflows.

Building the corpora and NLP infrastructure necessary for the automation of lexicographic tasks is lengthy process. In the meantime, there is no reason to fall back to entirely manually curated dictionary entries, which would only divert the lexicographers' precious language-specific expertise from the task of corpus development. There is no need to wait until a fully processed corpus and perfect NLP pipeline are in place, either. While the corpus is being developed, manually annotated sentences can be displayed to the public, without extra curation, via ad interim lexical resources through free and easy to set up tools such as Shiny or Lexonomy.

5. Acknowledgements

The project Lexicography in Motion is funded by the British Arts and Humanities Research Council. The Buddhist Translators Workbench was started with funding from the US National Endowment for the Humanities and is currently funded by the Mangalam

Research Center for Buddhist Languages.

6. References

Atkins, S. Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: OUP.

A Visual Dictionary of Buddhist Sanskrit. Accessed at: https://ligeialugli.shinyapps.io/VisualDictionaryOfBuddhistSanskrit/ (June15 2019).

Buddhist Translators Workbench. Accessed at https://btw.mangalamresearch.org/ (June 15 2019)

Chang, W., Cheng J., Allaire, J.J., Xie, Y., McPherson, J. (2018). *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny

Cook P., and Rundell, M., Lau J.L., Baldwin, T. (2014). Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples. In A. Abel, C. Vettori and N. Ralli (eds*.) Proceedings of the 16th EURALEX International Congress*. Bolzano: URAC research, pp. 319-328.

Edgerton, F. (1953). *Buddhist Hybrid Sanskrit* grammar and dictionary, 2 voll. New Haven: Yale University Press.

Gantar, P, Kosem, I., Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database, *International Journal of Lexicography* 29(2), pp. 200–225.

Garrett, E., Hill, N., Zadoks, A. (2014). A Rule-based Part-of-speech Tagger for Classical Tibetan. *Himalayan Linguistics*, (13)1, pp. 9-57.

Garrett, E., Hill. N, Kilgariff, A. , Vadlapudi, R., Zadoks. A. (2015). The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries. *Revue d'Etudes Tibétaines*, 32, pp. 51–86.

Garrett, E. (2017). Lexicography in Motion: Documentation. https://tibetan-nlp.github.io/lim-annodoc/

Gomez, L. O., Lugli, L. (2015). Buddhist Translators Workbench white paper. http://dx.doi.org/10.17613/M6866Z

Griffiths, P. (1981). Buddhist Hybrid English: Some notes on philology and hermeneutics for Buddhologists. *Journal of the International Association of Buddhist Studies* 4(2), pp. 17 -132.

Hackett, P. (2019). *A Tibetan Verb Lexicon*. Second edition. Boston: Snow Lion. First ed. 2003.

Handy, C. (2019). A context-free method for the computational analysis of Buddhist texts. In D. Veidlinger (ed.), *Digital Humanities and Buddhism: An Introduction*. Berlin: De Gruyter.

Hill, N. (2010). *A Lexicon of Tibetan Verb Stems as Reported by the Grammatical Tradition*. Munich: Bayerische Akademie der Wissenschaften.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal and J. DeCesaris (eds.), *Proceedings of the Thirteenth EURALEX International Congress*. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432.

Landau, S. I. 2001. Dictionaries: The art and craft of lexicography. Cambridge: CUP. First ed. 1984.

Lugli, L. (2015). Mapping meaning across time and cultures: innovations in Sanskrit lexicography. In *Words Dictionaries and Corpora: Proceedings of the 9th International Conference of ASIALEX.*

Lugli, L. (2018). Drifting in Timeless Polysemy: Problems of Chronology in Sanskrit Lexicography. *Dictionaries: Journal of the Dictionary Society of North America*, vol. 39(1), pp. 105–129.

Lugli, L. (2019). Words or terms? Models of terminology and the translation of Buddhist Sanskrit vocabulary. In Alice Collett (ed.) *Buddhism and Translation: Historical and Contextual Perspectives*, New York: SUNY.

Lugli.L. (In preparation). Towards Buddhist Sanskrit Corpus Linguistics: advances in segmentation, lemmatisation and syntactic inference for Buddhist Sanskrit.

Meelen, M., Hill, N. (2017). Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics*, 16(2), pp. 64–89.

Meelen, Marieke, Hill, Nathan, & Handy, Christopher. (2017). The Annotated Corpus of Classical Tibetan (ACTib), Part II - POS-tagged version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL [Data set]. Zenodo.

http://doi.org/10.5281/zenodo.822537.
Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In Electronic Lexicography in the 21st Century: Lexicography from Scratch. *Proceedings of the eLex 2017 conference*, Leiden.
Mianáin, P.O., Convery, C. (2014). From DANTE to Dictionary: The New English-Irish Dictionary. *Proceedings of the 16th EURALEX International Congress*, pp. 807-817.
Nasiruddin, M. (2013). A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. arXiv:1310.1425.
Pajzs, J. (2009). On the Possibility of Creating Multifunctional Lexicographical Databases. In H. Bergenholtz, S. Nielsen, & S. Tarp (eds.), *Lexicography at a crossroads. Dictionaries and encyclopedias today, lexicographical tools tomorrow*. Bern: Lang, pp. 327-354
Rundell, M., Atkins, S. (2011). The DANTE database: a User Guide. In M. Rundell & A. Killgarriff (eds.) *Proceedings of eLex 2011. Trojina: Institute for Applied Slovene Studies*, pp. 233–246.
Rundell, M. and A. Kilgarriff. (2011). 'Automating the creation of dictionaries: where will it all end?' In Meunier, F., G. Gilquin and M. Paquot (eds), *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: John Benjamins, pp. 257–282.