Dictionaries as collections of lexical data stories: an alternative post-editing model for historical corpus lexicography

Ligeia Lugli[1]

[1] SOAS, room 339, 10 Thornhaugh St, London WC1H 0XG, UK
E-mail: ll34@soas.ac.uk

Abstract

This paper proposes a model of dictionary post-editing inspired by data-journalism. It starts by problematising the parallel, drawn in the description of this year's eLex conference theme, between lexicographic and machine-translation post-editing. It then proceeds to outline data-journalism workflows and to illustrate how these may offer a suitable blueprint for automating and post-editing corpus-driven historical dictionaries of low-resource languages. In particular, the paper highlights the usefulness of adopting an iterative development model, whereby minimal automated entries are incrementally augmented with curated information, and of switching to data-visualisations as the main medium of communication.
Data-journalists concentrate much of their post-editing efforts in plotting the data into highly customised visualisations capable of narrating their interpretation of a story while also allowing multiple lines of inquiry. This paper suggests that historical lexicographers would benefit from similarly directing their post-editing efforts into weaving data into customised, lemma-specific, visualisations capable of guiding users towards further exploration.
The paper concludes with practical examples drawn from two ongoing historical dictionary projects, *A Visual Dictionary and Thesaurus of Buddhist Sanskrit* and *A Visual Dictionary of Tibetan Verb Valency*, which are adopting data-journalism workflows to post-edit automatically generated entries and data-visualisations into 'lexical data stories'.

Keywords: historical lexicography; data-journalism; post-editing; Sanskrit; Tibetan

1. Machine-translation post-editing for lexicography: a critique

For decades lexicography has been on a path of increasing automation. The late 90s and early 2000s vision of machines taking up the bulk of lexicographic work is now coalescing into reality (Grefenstette, 1998; Rundell, 2002). Hypothetical notions regarding the role of humans in a largely automated workflow are quickly being replaced by practical strategies for post-editing automated dictionary drafts. It is therefore a good moment to look at industries that already possess well established post-editing workflows and consider which could be most profitably adapted to which lexicographic endeavour.

The description of this year's eLex conference theme conceptualises lexicographic post-editing as akin to the post-editing practices honed in the field of machine-translation,[1] a parallel already drawn by Jakubíček a few years ago (Jakubíček, 2017). While machine-translation post-editing workflows may be profitably adapted to some

[1] '…This technological progress leads to new methodological approaches where most editorial work consists of post-editing of automatically created content – similarly to post-editing of machine-translated texts.' (eLex 2021 introductory paragraph, https://elex.link/elex2021/)

lexicographic projects (e.g. Baisa et al., 2019), they are not likely to constitute an optimal model for lexicography in general, and especially not for historical lexicography of low-resource languages, which is the focus of this paper. This is mostly due a fundamental difference in the nature and goals of translation and historical lexicography.

In machine-translation projects, computers generate a draft translation from an input text and humans refine it. The degree of manual refinement (i.e. post-editing) varies depending on how similar to a human-made translation the final product should be. 'Light' post-editing is often sufficient to ensure that the message of the source text is rendered accurately, while more labour intensive 'full' post-editing may be required to achieve a perfectly smooth reading experience in the target language, akin to a human translation (Nitzke et. al., 2019). In other words, machine-translation post-editing practices are articulated along two axes, accuracy, intended as faithfulness to the source text, and readability of the output text.

The relevance of these axes to historical lexicography is doubtful. While basic readability is indeed important, dictionary entries need not be specimen of great prose. Given their standardised wording and rigorously structured format, text generation templates should be capable of producing perfectly readable, if perhaps not enjoyable, dictionary entries (see Section 2 below). Post-editing for readability is therefore not likely to constitute a priority for many historical dictionary projects. Accuracy, by contrast, is a very likely priority. However, what constitutes accuracy in translation and in lexicography is entirely different. As such, machine-translation post-editing practices may well not be the best route to lexicographic accuracy.

The reason for this lies in a fundamental difference in the relationship between input data and output text in translation and lexicography. Translation aims at transforming its source data (by transposing it into another language), whereas lexicography aims at illustrating trends in its source data and deriving conclusions from them. This impacts the efficacy of text post-editing for accuracy in the two fields. In translation, manipulating the wording of the machine-generated draft directly affect its accuracy. Post-editing is thus an efficient path to improving the quality of computer-generated translations. While changing the text of automated dictionary drafts may also improve the overall dictionary quality, this is not an efficient path to increased accuracy. Lexicographic accuracy resides not so much in the wording of the entries as in the quality of sample, analysis and interpretation of the corpus data. Lexicographic accuracy is thus more directly impacted by addressing the representativeness of the corpus used, the level of detail of the linguistic annotation recorded in the corpus and the relevance of the statistical information automatically derived from it (Frankenberg-Garcia et al., 2020; Baisa et al., 2019). As it will be discussed in section 3, post-editing may not be the most efficient way to address these matters in historical lexicography of low resource languages, where the efforts could rather be concentrated in enriching a small corpus with detailed linguistic information.

Moreover, what constitutes an accurate representation of the input data is much more subjective in historical lexicography than it is in machine-translation. Interpretation is typically straightforward in automatically translated texts—literary works, puns and ambiguous prose lying still largely beyond the scope of machine-translation, and best translated from scratch by humans (Nitzke et. al., 2019). This means that machine-translation post-editing can realistically aim to achieve an uncontroversial version of the translated text; a version that is going to be equally useful to all its readers.

The situation is more complex in lexicography. Much of what goes into a dictionary entry, from sense categorisation and sense descriptions up to example selection, is highly interpretive. In the case of historical lexicography, matters of philological uncertainty, disputed dating and difficulties of interpretation further complicate the picture. Adopting a machine-translation post-editing model in historical lexicography hardly does justice to this complexity, or to dictionary users. It implies a conceptualisation of dictionary entries as a definitive top-down account of a word's semantics and usage, which risks misrepresenting interpretation and subjective choices as purely descriptive accounts. This vastly limits the usefulness of historical dictionaries as tools for research. Post-editing models that allow users to pursue different interpretations of the data and provide a transparent record of lexicographers' editing choices may yield more versatile and useful resources.

Finally, a post-editing model inspired by machine-translation raises concerns of sustainability for historical dictionary projects that depend on public funding. Public funding cycles for humanities projects are relatively short, covering typically a period of three years in the UK and USA (e.g. schemes funded by the Arts and Humanities Research Council and National Endowment for the Humanities). As a result, historical dictionary projects often need to produce a minimally viable product very quickly in order to showcase their outputs early and secure follow-up funding for further work. If dealing with low-resource languages or specialised domains, they also often need to create and process corpora from scratch and thus invest a significant portion of their first funded period into developing the source data necessary for their dictionaries. Under these circumstances, it is advisable to develop dictionaries iteratively, by first publishing automated entries based on corpus data and then gradually refining and augmenting them through further iterations (Lugli, 2019). This makes it possible to align lexicographic outputs with funding cycles, but it is important to note that this model is efficient only in so far as there is no overlap in the work required for each iteration. It is doubtful that this is best achieved through the adoption of post-editing practices inspired by machine-translation.

In machine-translation contexts, the choice between different levels of post-editing (bare machine output, light post-editing or full post-editing) occurs early on in a project. The literature on automated translation construes the relationship between light post-editing and full post-editing as one of alternative editorial strategies, rather than as a progression between different editorial stages, since arguably both involve

much of the same tasks (see the post-editing decision tree in Nitzke et al. 2019, 246). While the practices developed for machine-translation can surely be adapted to the needs of lexicography (as accomplished, for example in the project described in Baisa et al., 2019), in light of the limitations outlined above it seems useful to expand the pool of reference models available for dictionary post-editing. I propose that we consider one model that is remarkably close to historical lexicography in several respects: data-journalism.

2. Text automation and post-editing in journalism

Data-journalism is a branch of journalism that focusses on deriving news stories from datasets and typically conveys much of the information through data-visualisations.[2] The complexity of data-journalism pieces ranges from relatively simple graphs and narratives, such as those charting the spread of COVID-19, ubiquitous in newspapers these days, to the more nuanced and interpretive pieces published in dedicated data-journalism outlets, such as *The Economist*'s *Graphic Detail*.

Like translation, journalism has undergone considerable levels of automation in recent years. As with machine-translation, drafts of news pieces are now routinely generated automatically and then refined through human curation (Marconi, 2020; Diakolpoulos, 2019; Graefe, 2016). The processes of text generation and post-editing, however, differ between the translation and news industries. The difference is, again, rooted in the relationship between input data and automatically generated output. While translations transpose the input data into a new language, news pieces elaborate on the input data, typically producing entirely new text from and about numeric inputs.

An output text's relationship with the input data varies depending on the type of news. Reports on sport matches or election results summarise the input data; financial news may highlight trends and changes in assets' value; in-depth analyses may draw conclusions from the input data, or use them to support a specific argument. While all kinds of data-based news can be (and indeed are) automated, the degree and quality of the automation, as well as the post-editing strategies required to reach a publishable product vary.

There is consensus in the literature on automated journalism that the best automated output is achieved with types of news that have a relatively rigid format, a predictable vocabulary, rely on highly structured data and describe (rather than interpret) the input data. These types of news include market and weather reports as well as sports

[2] My use of the term data-visualisation is close to the definition provided by Bakakis: 'Data visualization is the presentation of data in a pictorial or graphical format, and a data visualization tool is the software that generates this presentation. Data visualization provides users with intuitive means to interactively explore and analyze data, enabling them to effectively identify interesting patterns, infer correlations and causalities, and supports sense-making activities. ' (Bakakis, 2018), but I extend my application of the term to cover cases of static (i.e. non-interactive) data-visualisation as well.

and election results—all of which have been routinely automated for years (Carlson, 2015; Diakopoulos, 2019). For these types of news, automated text is published with minimal or no human post-editing (Graefe et al., 2018; Diakopoulos, 2019).[3] This is not to say that these news pieces do not require any human labour at all. Rather, the labour is concentrated in pre-processing. Before any automated news writing can take place, humans need to prepare the data and templates that will be used to generate the text of news pieces. Data preparation includes the usual steps of cleaning, wrangling and verification, and needs to be performed on any new data used. This appears to be the weakest link in run-of-the-mill news automation, as the errors discussed in the literature are all due to poorly pre-processed data (e.g. Diakopoulos, 2019: 133; Marconi, 2020: 69). Template preparation is more robust, but rapidly evolving. Traditionally, templates for automatic text-generation are 'hard coded'. News editors prepare set templates for each type of news, detailing the order in which the information is to be presented, as well as alternative sentence structures to be used convey each piece of information and pools of synonyms to choose from to ensure some variation in the automated texts. The results of this procedure are consistently good and often indistinguishable from human writing (Diakopoulos, 2019: 126). In recent years, the creation of templates has been partially automated and machines are now able to structure a piece and concatenate (and in some cases craft) sentences on the basis of rules and/or statistical models derived from news corpora (Diakopoulos, 2019: 98 ff.; Leppänen et al., 2017). This obviously leads to faster pre-processing by drastically reducing the need for detailing domain-specific templates. The overall time and labour required to achieve a publishable product, however, is not reduced. Dynamically created templates tend to introduce problems of readability and thus require more post-editing efforts. Unsurprisingly, the news industry prefers to invest resources in labour-intensive template creation and dispense with (or minimise) post-editing, rather than opt for the reverse (Diakopoulos, 2019). This is an efficient choice as even though they may not generalise well across different types of news, detailed templates are still re-usable for all news within a given category. Post-editing by contrast is piece-specific; it is not re-usable at all, at least for now.[4]

The opposite is true for news stories that are based on data but require investigation, interpretation and are best conveyed through original narratives. That is, the type of news stories that is most typically referred to as 'data-journalism'.[5] Even though data

[3] RADAR, a leading news project, only manually checks the output of one in ten automated news pieces (Diakopoulos 2019, 134).

[4] See Diakolopoulos's brief discussion of 'distant editing' as a prominent desideratum in the news industry (Diakopoulos 2019, 134 and 247-248).

[5] Several definitions of data-journalism and discussions of its relative position within the field journalism vis-à-vis other computer-enhanced forms of news-making have been put forward (see Coddington 2018 for a comprehensive review). For the purposes of this paper, the generic characterisation of data-journalism as an approach to crafting news stories that is centred on the acquisition, analysis, interpretation and publication of data will suffice (cf. Usher, 2016: 90; Howard, 2014: 2-5; both cited in Coddington, 2018: 17).

play a central role in these stories, they cannot simply be plugged in a text template. The narrative is too unique to be amenable to templates; no matter how sophisticatedly constructed they might be (Stray, 2019; Caswell and Doerr, 2018). The efficient choice for these stories is to switch from a paradigm of pure automation to one of augmentation, whereby machines generate a minimal description from the data and leave it to journalists to investigate and flesh out the narrative of the story (Diakopoulos, 2019: 46ff; Graefe, 2016: 29). While the journalism literature refers to this process as 'augmentation' or 'human-machine interaction' (Marconi, 2020: 69-71; Diakopoulos, 2019: 247-248), it is a form of post-editing, in so far as it amounts to the manual curation of an automatically generated draft. Still, it differs from machine-translation post-editing in two important respects: it is iterative and not centred on text.

In data-journalism, the initial automated summary of data can constitute a minimal viable product (or 'minimally viable story', Marconi, 2020). This product may not be fit for publication in a newspaper, but it is usually good enough to be immediately released in the form of a blog post or as a news alert (Young and Hermida, 2015). The automated summary can then be enriched with more information and interpretation in successive stages—possibly depending on the amount of interest that each iteration of the story generates among the public (Marconi, 2020).[6] Besides being efficient for news production, this iterative story development is also empowering for the reader. It provides early and comprehensive access to granular data that would otherwise not be available, such as real time information on local crime or a detailed breakdown of minor election results, which journalists would rarely have the time to report manually (Young and Hermida, 2015; Leppänen et al., 2017; Marconi 2020).

Unredacted automatic reports may not make for a very enjoyable read, though. Fortunately, the dullness of automated text can be entirely bypassed by presenting the automated data summary in the form of data-visualisations. Reliance on data-visualisations is one of the most salient features of data-journalism (Coddington, 2018; Kennedy et al., 2019).[7] Tools for the automatic identification of potentially newsworthy leads typically supply journalists with visual analytics (Diakopoulos, 2019: 57, 48ff; Wiedmann, 2018; Stray 2019), and systems are in place to automatically generate publication-ready data-visualisations to accompany data-driven news (Alhalaseh et al., 2018). The initial automatically generated minimally viable story could thus take the form of a graph or data-visualisation dashboard (e.g. Diakopoulos, 2019: 49 fig 2.1).

Post-editing also focusses on visualisations. Much of the educational literature on how to craft data-journalism stories stresses the importance of editing the visualisations

[6] See Marconi 2020, chapter 3 for a detailed explanation of iterative journalism.

[7] Data-visualisations are perceived by some as having replaced writing as the "main semiotic mode" of journalistic storytelling (Kennedy et al., 2019).

accompanying the story so that they communicate the main points of the narrative, highlight the author's interpretation of the data and guide the user towards specific insights (Thudt et al., 2018; Stopler et al., 2018; Kennedy et al., 2019). Given the interpretative nature of data-journalism, another topic that is emphasised in this literature is the role of interactive data-visualisation in encouraging users to explore multiple lines of inquiry, reach different interpretations and reveal bias (Thudt et al., 2018; Diakopouls, 2018, 246). By curating data-visualisations and letting users explore the dataset used for a story, journalists increase transparency and civic engagement, two cornerstones of data-journalism ethics (Coddington, 2018; cf. Kennedy et al., 2019). These practices may also help historical lexicographers meet the needs of their audiences.

3. A data-journalism post-editing model for lexicography

The post-editing practices developed for data-based news pieces could be profitably transferred to historical corpus lexicography of low-resource languages. This subset of lexicography possesses some characteristics that make it an especially good fit for the newsroom's approach to post-editing.

First of all, its low-resource aspect. Limited budget and manpower make it necessary to prioritise efforts very carefully, and dependence on public funding makes iterative dictionary development especially suitable for this type of lexicography. Under these circumstances, the newsroom practice of shifting labour from post-editing single-purpose texts to preparing data and templates for the automatic generation of multiple texts is appealing.

This model of labour allocation may even work better in lexicography than in news production, for two reasons. As mentioned earlier, poorly prepared data and complex narratives are the two main obstacles to post-editing-free news automation. Neither of these apply to lexicography. Data preparation is challenging in journalism because news data change continuously and thus require constant monitoring and checking. By contrast, the data used for historical dictionaries typically amounts to a language corpus that only needs be prepared once. Moreover, while only a fraction of news stories fit the requirements for template-based text generation, dictionary entries, with their fixed structure, formulaic phraseology and well-ordered integration of corpus data, are perfectly amenable to simple templates, which can easily be enriched with dynamic data-visualizations to allow users to actively engage with the data behind the entries. Indeed, the dictionary post-editing model inspired by machine-translation also leverages this characteristic of lexicographic entries by slotting automatically extracted and sorted corpus data in specified fields within an entry (e.g. Měchura, 2017). The difference in the data-journalism model is that an automated minimally viable entry can be published without any post-editing and still be highly engaging thanks to reliance on interactive data-visualisations and highly curated corpus data (cf. Baisa et al., 2019).

This, again, works best for low-resource historical languages. For three reasons. First, the corpora available for these languages are typically rather small by contemporary standards and are often created for specific lexicographic purposes. This allows for more fine-grained annotations to be encoded in the corpus than is typically possible for larger corpora. It also allows for manual curation of the annotations, which, as a result, may be more accurate and detailed than the automated tagging typical of large corpora (Lugli, 2019). Such accurate and fine-grained annotations in turn allow for a wider range of information to be automatically derived from the corpus and plugged into entry templates, thus enabling the creation of fairly rich automated entries (see the next section for examples).

Second, new historical dictionaries of low-resource languages typically bring to the public lexical data that would not otherwise be available (e.g. data from newly created corpora or newly discovered manuscripts). Hence their audiences are likely to benefit from early access to new lexicographic material, even if it is in the minimal form of an automated entry.

Finally, historical dictionaries of low-resource languages tend to be used for research purposes, often by highly trained academics. Some of the work typically required in dictionary post-editing, such as checking the automated selection of examples, can therefore be offloaded to users, who may even prefer to filter through examples themselves, using custom parameters, rather than be given a fixed set of sentences pre-selected by lexicographers.[8] Given the uncertainty surrounding much historical material, especially for low-resource languages, these users are also likely to prefer having the option of engaging directly with the data rather than being given solely a top-down interpretation of the meaning and evolution of a given lemma. A purely automated entry presenting annotated corpus data could thus serve this user pool, especially if it offers the possibility to explore the data interactively.

To this end, the data-journalism practice of publishing automated news stories as, or with, data-visualisations is, again, better suited for historical lexicography than the machine-translations model of a text-centred dictionary entry. Since dictionaries have been moving away from prescriptivist definitions and towards descriptions of words' use, conveying the content of lexicographic entries through data-visualisation has become easier. An automated description of corpus information is easier to render graphically than verbally. Easily programmable data-visualisations can efficiently represent data that would require complex sentences and elaborate text-generation templates to be described in text (see next section for examples).

Overall, data-visualisations require less post-editing than text. Problems of syntax, infelicitous wording or clumsy sentence concatenation do not apply to charts. Still,

[8] This is the feedback we received from prospective users of both the Tibetan and Sanskrit dictionaries.

post-editing data-visualisation is advisable. Automatically generated charts may fail to highlight the most interesting aspect of the data, or obfuscate important patterns in a sea of data points. Especially so, if the same set of visualisations is applied to all lemmata in the dictionary, regardless of their semantic characteristics or distributional patterns. Some charts will inevitably fit better one lemma than another. Hand picking the best chart for each lemma and selecting the most effective colour scheme or interactive options for each type of information is thus an important task.

Data-visualisation post-editing can be the focus of a second iteration of the dictionary. Here the automated entries generated in the first iteration can be augmented with a view to guide users through the data. A third iteration can further augment the entries with the addition of a text narrative that explains the lexicographer's interpretation of the data. This final iteration would combine high-level lexicographic curation with interactive data-exploration, thus balancing guided and self-directed use of the resource and allowing multiple interpretations.

Given that post-editing is both labour intensive and single-purpose, it may be expedient to limit it to a subset of entries (cf. Baisa et al., 2019). Following the data-journalism model, lexicographers could concentrate their manual efforts on lemmata that are deemed especially interesting, either because they attract the most views from users or because they satisfy some predefined statistical test. For example, polysemic words that display dramatic diachronic changes could be the focus of detailed entries that explain their development and semantic plasticity, while monosemic words that are homogeneously distributed across periods may be satisfactorily represented by automated minimal entries.

4. Examples from Tibetan and Sanskrit lexicography

A post-editing model inspired by data-journalism has been applied to two historical dictionaries of low-resource languages currently under development. Both dictionaries are still undergoing their first iteration and are presently best characterised as working prototypes. Both are highly specialised lexical resources, one is a dictionary and thesaurus of Buddhist Sanskrit aimed at translators of Buddhist literature (*A Visual Dictionary and Thesaurus of Buddhist Sanskrit*), the other is a diachronic valency lexicon of Tibetan verbs (*A Visual Dictionary of Tibetan Verb Valency*).

The two resources differ completely in content and aim, but have been developed following the same 'deferred post-editing' iterative model, whereby a completely automated version of the dictionary is released as proof of concept and is then followed by incrementally augmented iterations (Lugli, 2019). While both dictionaries have adopted an approach to post-editing that closely resembles that of automated journalism (especially the one described in Marconi, 2020), it should be noted that this connection is made *a posteriori* and the dictionary development model was not originally inspired by data-journalism. By contrast, new editorial directions regarding data-visualisation post-editing, which will be introduced shortly, have been explicitly

modelled after data-journalism best practices.

As with automated news, the dictionary projects discussed here shifted the bulk of lexicographic work from post-editing to data and template preparation. For the Tibetan project, most of the lexicographic work consisted of annotating a small diachronic corpus with verb argument structure, which is the primary focus of this resource. Additionally, a sample of about five thousand sentences instantiating the top hundred most frequent verbs are also being annotated semantically, by specifying the meaning that the headword verb takes in each sentence. Since other good dictionaries of Tibetan verbs exist, in this project we have opted for using the sense categorisation provided in pre-existent resources (specifically Hill, 2010). By contrast, the Buddhist Sanskrit project focusses on fine-grained semantic analysis, with lexicographers annotating a small corpus of sampled sentences with original information regarding word senses, semantic prosody, as well as conceptual and syntactic relations. The annotation process in both projects has been time consuming, but has resulted in a re-usable, multi-purpose dataset that makes lexicographic analysis not only time efficient but also completely transparent by clearly associating each data point with the corresponding interpretation provided by the lexicographers (Lugli, 2019).[9]

For both dictionaries, the annotated corpus data is plugged into a programmatic template that for each headword generates three main types of outputs: 1) a short text summary, 2) a variety of interactive data-visualisations and 3) a dynamic list of examples that can be filtered according to various parameters. At present, the automated text is minimal. In the Sanskrit dictionary it only provides a breakdown of the headword's senses, whereas in the Tibetan valency lexicon it also adds a summary of frequency, diachronic distribution and valency structure of each headword and, where applicable, the light verb constructions (a type of multiword expression) in which it participates (Figure 1). Following feedback from peers and users, we will expand the automated text summaries to include more descriptive prose and some examples.

Examples are currently displayed in a separate tab in both resources. The Sanskrit dictionary has a 'quick examples' tab that displays examples that have been manually selected by lexicographers during corpus annotations and a 'more examples' tab that allows users to access all the sentences that have been annotated for a lemma and filter them by genre, sense, semantic prosody and grammatical features. The Tibetan lexicon does not offer hand-picked examples. Instead, it sorts the annotated sentences according to a 'good example' score inspired by the Sketch Engine's Gdex paradigm (Kilgariff et al., 2008). Soon, it will offer users the possibility to manipulate this score according to their own preferred parameters. After all, what constitutes a good example largely depends on what a user wishes to see exemplified. The default score aims at prioritising sentences that express a complete thought, are relatively short, contain no anaphoric references and only minimal 'noise', such as long lists of verbs

[9] Lugli 2019 discusses in detail the efficiency of this workflow.

and strings of modifiers (Lugli, 2019, 208-209). Yet, some users may wish to see examples of the interplay between valency patterns and anaphoric markers, or see how the headword verb is strung together with other verbs in formulaic lists.

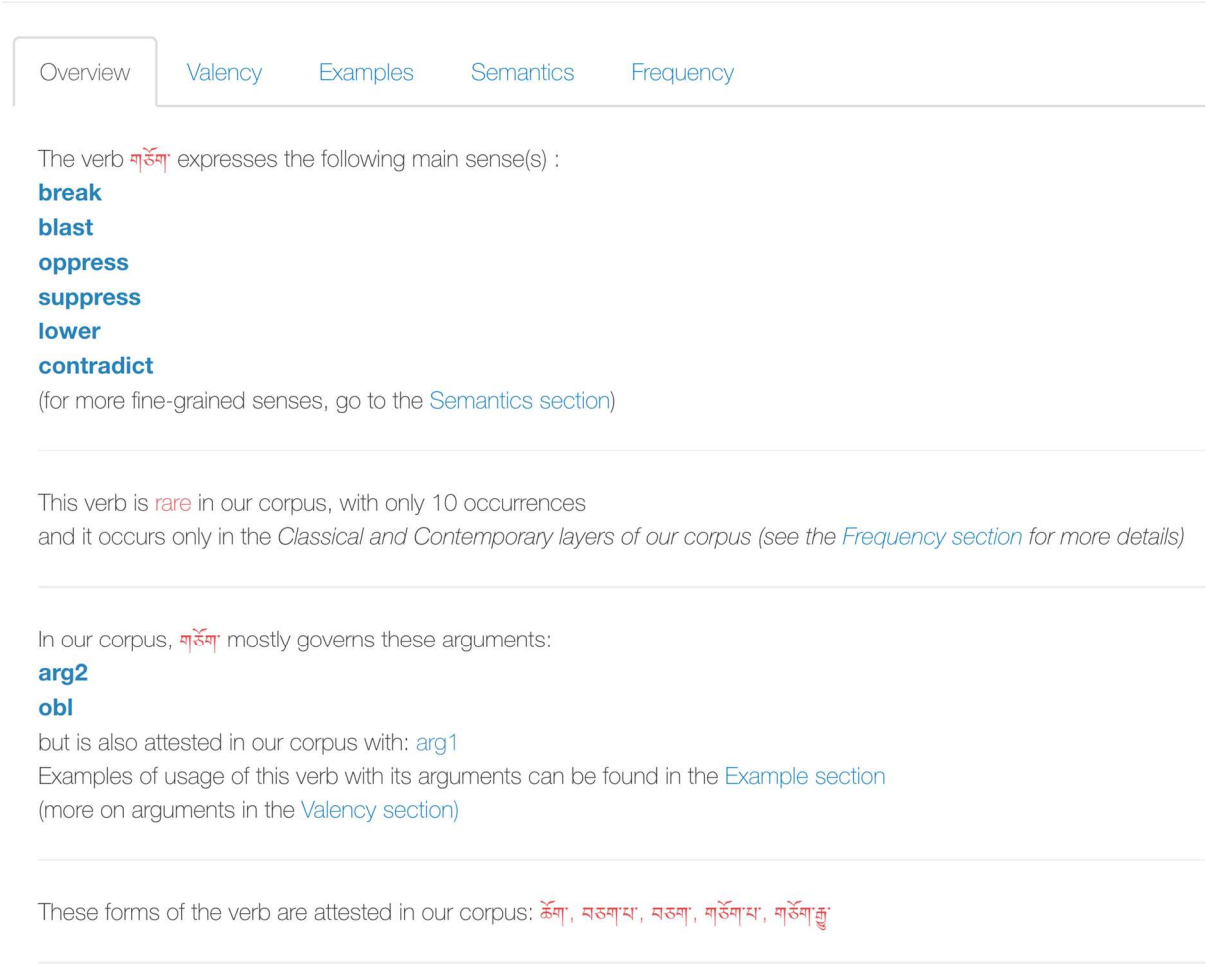

Figure 1. Automatically generated lemma overview in the Visual Dictionary of Tibetan Verb Valency (mangalamresearch. shinyapps.io/VisualDictionaryOfTibetanVerbValency/, accessed on 8/4/2021)

Examples are not the only area where the display preferred by lexicographers and users may differ. Both dictionaries allow users to interact with the graphs to view their own preferred combination of variables, switch between different types of charts, or change between normalised and absolute frequencies. More importantly, the Sanskrit dictionary allows users to customise periodisation and other metadata and adjust all data-visualisations accordingly. This is crucial when dealing with Sanskrit literature, where dating of texts is uncertain and often hotly disputed (Lugli, 2018).

In sum, the first iteration of both dictionaries offers users a wealth of manually annotated data and the possibility to explore it interactively and, potentially, to reach their own conclusions. One important limitation of these first iterations is that they

do not make explicit the conclusions reached by the lexicographic teams. While the interpretation of each sentence is granularly recorded in the source data in the form of linguistic annotations, the automated entries do not provide an overall interpretation of the semantic or syntactic history of headwords. Such interpretation is the object of our post-editing phase and constitutes the focus of further iterations.

A second iteration is currently being planned for the Buddhist Sanskrit dictionary. It will centre around the creation of 'lexical portraits', curated interpretations of the data that mix narrative and edited data-visualisations. While the details of our post-editing pipeline are still being tried and tested, the general principles are clear. They are inspired by data-journalism workflows in that they aim to lead lexicographers and users alike through a progression from a minimal automated summary to an interpretive explanation that blends human-written text with purpose-specific graphs. The process starts with lexicographers receiving automated summaries for each lemmata. These summaries take the form of visual analytics and touch upon four main areas, 1-2) lemma and sense distribution over subcorpora, 3) lexical context in which each word-sense tends to occur and 4) distribution of the semantic prosody of each sense over the subcorpora. The automated dashboard also highlights the cross-section of subcorpora where the most change is detected (e.g. periods or genre or philosophical tradition). Lexicographers create a text narrative that explicitly interprets the information provided in the automated summary and relates it to areas of interest for translators (the primary target audience of this dictionary), such as register, level of technical specialisation, connotation and comparison with near-synonyms. While drafting the narrative, lexicographers are asked to 1) refer to specific example sentences (examples are taken from the first iteration of the dictionary), 2) select the appropriate chart to illustrate each aspect of the data that is discussed in the narrative and 3) edit the charts to maximise their communicative power. This last point is probably the least practiced in historical lexicography, so a few examples are in order.[10]

The following examples are taken from an entry prototype that we are developing for the second iteration of the Buddhist Sanskrit dictionary. Since we are still working on this prototype, only a single proof-of-concept entry is currently available online in this new format, the lexical portrait of the word *vitarka*. The prototype is accessible from the dictionary entry on *vitarka*, but this is presently not yet integrated in the dictionary application, but hosted separately at mangalamresearch.shinyapps.io/LexicalPortrait_Vitarka/.

Most of the charts in this prototype are post-edited versions of the automated charts included in the automated summary. An automated graph showing the frequency of the lemma compared to its near synonyms, for example, has been edited by manually trimming the pool of near-synonyms shown to enhance the readability of the graph.

[10] This is not to say that no efforts have been made in the direction of data-visualisation post-editing within historical lexicography, but these efforts still seem very rare (e.g. Hoenen, 2018).

Other charts have been edited to facilitate interpretation. For example, the automated summary contained a barchart illustrating the normalised frequency of the lemma in each text type. It was clear from the chart that the lemma is dramatically more frequent in the genre *śāstra* (treatise) than in the other genres, but the fragmentation into several bars obfuscated the focal contrast between the frequency of *vitarka* in *śāstras* and in the rest of Buddhist literature, which is explicitly referred to in the narrative accompanying the graph. To make the comparison clearer, we added a second chart that displays the cumulative frequency of the headword in all other genres (figure 2). Finally, we edited a wordcloud to highlight the link between lexical context and semantics. The automated version of this wordcloud highlighted the words that surround *vitarka* according to their collocational strength. Some rather obscure words that happen to co-occur with *vitarka* with statistically significant frequency were prominently displayed. The resulting data-visualisation was not very informative, as the highlighted words scarcely contributed to the interpretation of the headword's semantics. To improve on this, we manually experimented with different parameters and eventually changed them to highlight words according to the number of texts in which they co-occur with the headword. This produced a more informative picture where the most prominent items clearly point to the two different senses of the headwords.

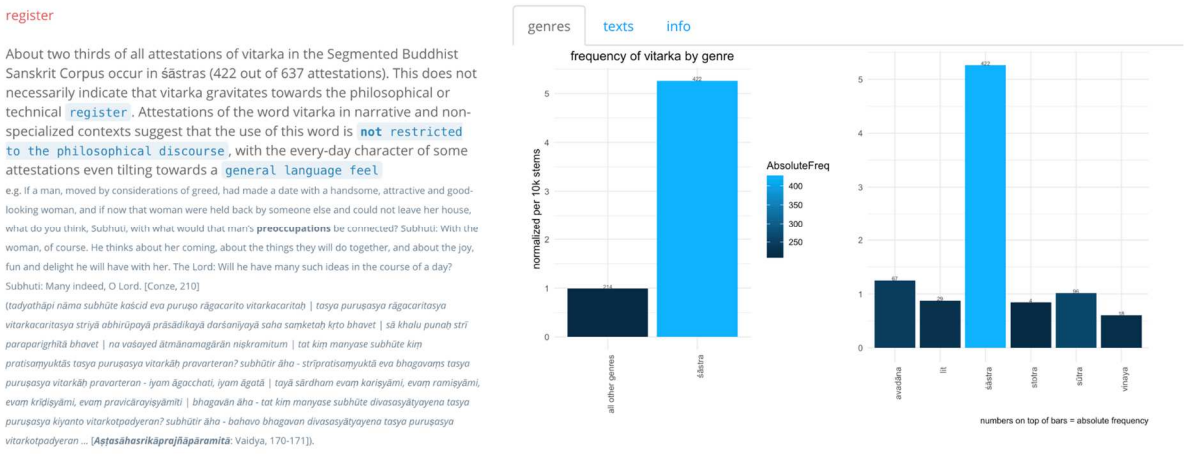

Figure 2. A portion of the prototype lexical portrait (mangalamresearch.shinyapps.io/LexicalPortrait_Vitarka/, accessed on 8/4/2021).

The parameters used to generate each graph are detailed in the 'info' tab accompanying each graph. The text slotted in the 'info' sections is automatically generated via a template that describes the default parameters used to generate the graph and is edited whenever the parameters are manually changed. The text of the narrative, by contrast, is unlikely to be amenable to automation. While we may experiment with generating an automatic draft for the lexicographers to post-edit in a machine-translation fashion, it seems that the interpretive and original content of the narrative is better suited to the augmentation model of data-journalism, whereby only the visual analytics are automatically generated and the storytelling is left to the human author.

5. Acknowledgements

Part of the work presented in this paper has been funded by UKRI as part of the project *Lexicography in Motion: A History of the Tibetan* Verb (AH/P004644/1). The Sanskrit dictionary is funded by the Mangalam Research Center for Buddhist Languages.

6. References

A Visual Dictionary and Thesaurus of Buddhist Sanskrit. Accessed at: mangalamresearch.shinyapps.io/VisualDictionaryOfBuddhistSanskrit/ (30 March 2021)

A Visual Dictionary of Tibetan Verb Valency. Accessed at: mangalamresearch.shinyapps.io/VisualDictionaryOfTibetanVerbValency/ (30 March 2021)

Alhalaseh, R., Munezero,M., Leinonen, L. & Leppänen, L. (2018). Towards Data-Driven Generation of Visualizations for Automatically Generated News Articles. *Proceedings of the 22nd International Academic Mindtrek Conference.*

Baisa, V. et al. (2019). Automating Dictionary Production: a Tagalog-English-Korean Dictionary from Scratch. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius C. (eds.) *Smart Lexicography: eLex 2019*, pp. 805–818.

Bakakis, N. (2018). Big data visualization tools. *arXiv* 1801.08336.

Caswell, D. & Doerr, K. (2018). Automated Journalism 2.0: Event-Driven Narratives, from Simple Descriptions to Real Stories. *Journalism Practice*, 12(4), pp. 477–496.

Coddington M. (2018). Defining and Mapping Data Journalism and Computational Journalism: A Review of Typologies and Themes. In S. Eldridge II (ed.) *The Routledge Handbook of Developments in Digital Journalism Studies*. New York: Routledge.

Diakopoulos, N. (2018). Ethics in Data-Driven Storytelling. In N. Henry Riche et al. (eds.) *Data-Driven Storytelling*. London: CRC, pp. 233–247.

Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Frankenberg-Garcia, A., Rees, G. P. & Lew, R. (2020). Slipping through the Cracks in e-Lexicography. *International Journal of Lexicography*, doi: 10.1093/ijl/ecaa022.

Graefe, A. (2016). *Guide to Automated Journalism.* Columbia University, Tow Center for Digital Journalism.

Graefe, A., Haim, M. & Brosius, H. B. (2018). Perception of Automated Computer-Generated News: Credibility, Expertise and Readability. *Journalism*, 19(5), pp. 95–610.

Graphic Detail. Accessed at: https://www.economist.com/graphic-detail (28 May 2021)

Grefenstette, G. (1998). The Future of Linguistics and Lexicographers: will there be Lexicographers in the Year 3000? In T. Fontenelle et al. (eds.) *Proceedings of the Eighth Euralex Conference*, Liège.

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

Hoenen, A. (2018). Annotated Timelines and Stacked Area Plots for Visualization in Lexicography. *Elexis Workshop*, Galway 2018.

Jakubíček, M. (2017). The advent of post-editing lexicography. *Kernerman Dictionary News*, July 2017.

Kennedy, H. et al. (2019). Data Visualisations: Newsroom Trends and Everyday Engagements. In J. Gray & L. Bounegru (eds.) *The Data Journalism Handbook 2: Towards a Critical Data Practice*. Amsterdam: Amsterdam University Press.

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

Leppänen, L., Munezero, M., Granroth-Wilding, M., Toivonen, H. (2017). Data-Driven News Generation for Automated Journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 188–197.

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). Smart lexicography for under-resourced languages: lessons learned from Sanskrit and Tibetan. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius C. (eds.) *Smart Lexicography: eLex 2019*, pp. 198–212.

Marconi, F. (2020). *Newsmakers: Artificial Intelligence and the Future of Journalism*. New York: Columbia University Press.

Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) Electronic Lexicography in the 21st Century: Lexicography from Scratch. *Proceedings of the eLex 2017 conference*, Leiden.

Nitzke, J., Hansen-Schirra, S., Canfora, C. (2019). Risk management and post-editing competence. *The Journal of Specialised Translation*, 31, pp. 240–259.

Rundell, M. (2002). Good Old-fashioned Lexicography: Human Judgement and the Limits of Automation. In M.-H. Corréard (ed.) *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins*. Euralex.

Stopler, Ch.D., Lee, B., Henry Riche, N. & Statsko, J. (2018). Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Visual Stories. In N. Henry Riche et al. (eds.), *Data-Driven Storytelling*. London: CRC, pp. 85–105.

Stray, J. (2019). Making Artificial Intelligence work for Investigative Journalism. *Digital Journalism*, 7(8), pp. 1076–1097.

Thudt A, Walny J., Gschwandtner, Th., Dykes, J. & Statsko, J. (2018). Exploration and Explanation in Data-Driven Storytelling. In Nathalie Henry Riche et al. (eds.), *Data-Driven Storytelling*. London: CRC, pp. 59–83.

Wiedmann, G., Yimam, S. M. & Biemann, Ch. (2018). A Multilingual Information Extraction Pipeline for Investigative Journalism. *arXiv* 1809.0022v.1

Young, M. L. & Hermida, A. (2015). From Mr. and Mrs. Outlier to Central Tendencies. *Digital Journalism* 3(3), pp. 381–397.