Many a Little Makes a Mickle – Infrastructure Component Reuse for a Massively Multilingual Linguistic Study

Lars Borin
University of Gothenburg
Sweden

Shafqat Mumtaz Virk
University of Gothenburg
Sweden

Anju Saxena
Uppsala University
Sweden

lars.borin@svenska.gu.se, virk.shafqat@gmail.com, anju.saxena@lingfil.uu.se

Abstract

We present ongoing work aiming at turning the linguistic material available in Griersons classical *Linguistic Survey of India* (LSI) into a digital language resource, a database suitable for a broad array of linguistic investigations of the languages of South Asia and studies relating to language typology and contact linguistics. The project has two additional main aims: (1) to conduct a linguistic investigation of the claim that South Asia constitutes a linguistic area; (2) to develop state-of-the-art language technology for automatically extracting the relevant information from the text of the LSI. In this presentation we focus on how, in the first part of the project, a number of existing research infrastructure components have been 'recycled' in order to allow the linguists involved in the project to quickly orient themselves in the vast LSI material, and to be able to provide input to the language technologists designing the information extraction from the descriptive grammars.

1 Introduction: South Asian Linguistics and the *Linguistic Survey of India*

South Asia (also "India[n subcontinent]") with its rich and diverse linguistic tapestry of hundreds of languages, including many from four major language families (Indo-European>Indo-Aryan, Dravidian, Austroasiatic and Tibeto-Burman), and a long history of intensive language contact, provides rich empirical data for studies of linguistic genealogy, linguistic typology, and language contact.

South Asia is often referred to as a *linguistic area*, a region where, due to close contact and widespread multilingualism, languages have influenced one another to the extent that both related and unrelated languages are more similar on many linguistic levels than we would expect. However, with some rare exceptions (e.g., Masica, 1976) most studies are largely impressionistic, drawing examples from a few languages (Ebert, 2006).

In this paper we present our ongoing work aiming at turning the linguistic material available in Grierson's classical *Linguistic Survey of India* (LSI; Grierson, 1903–1927) into a digital language resource, a database suitable for a broad array of linguistic investigations of the languages of South Asia.

The LSI presents a comprehensive survey of the languages spoken in South Asia conducted in the late nineteenth and the early twentieth century by the British government. Under the supervision of George A. Grierson, the survey resulted into a detailed report comprising 19 tomes, some 9500 pages in total. The survey covered 723 linguistic varieties representing major language families and some unclassified languages, of almost the whole of nineteenth-century British-controlled India (modern Pakistan, India, Bangladesh, and parts of Burma). For each major variety it provides (1) a grammatical sketch (including a description of the sound system); (2) a core word list; and (3) text specimens (including a morpheme-glossed translation of the *Parable of the Prodigal Son*). The linguistic sketches include information on some of the features that have been used in defining South Asia as a linguistic area, e.g. retroflexion, reduplication, compound verbs, word order, but goes considerably beyond these, offering the possibility of a broad comparative study of South Asian languages.

The project where this work is conducted has two main goals: (1) to conduct a linguistic investigation of the claim that South Asia constitutes a linguistic area; (2) to develop state-of-the-art language technology, in the form of an information extraction (IE) method for automatic retrieval of linguistic features from the LSI text.

The full text of the LSI (only the Latin-script portions) was digitized by a commercial digitization service using double keying, which has resulted in a digital version of very high quality. The amount of text that has been digitized so far is well in excess of one million words.

In the first phase of the project, the linguists in the project team needed to quickly orient themselves in the vast material of the LSI, both so that they would get an overview of the linguistic features present in the descriptive grammars, and to be able to provide input to the language technologists designing the IE application. In particular, we require gold-standard data on which we can evaluate our IE experiments. This dataset is prepared using a standard methodological tool in large-scale comparative linguistics, viz. the linguistic questionnaire. In our case, the questionnaires contain mostly yes–no questions – e.g., "Does the language mark dual in at least one personal pronoun?" – and, inevitably, some dependencies among questions, e.g., if the answer to the pronominal dual question is "yes", there are follow-up questions about first, second and third person pronouns.

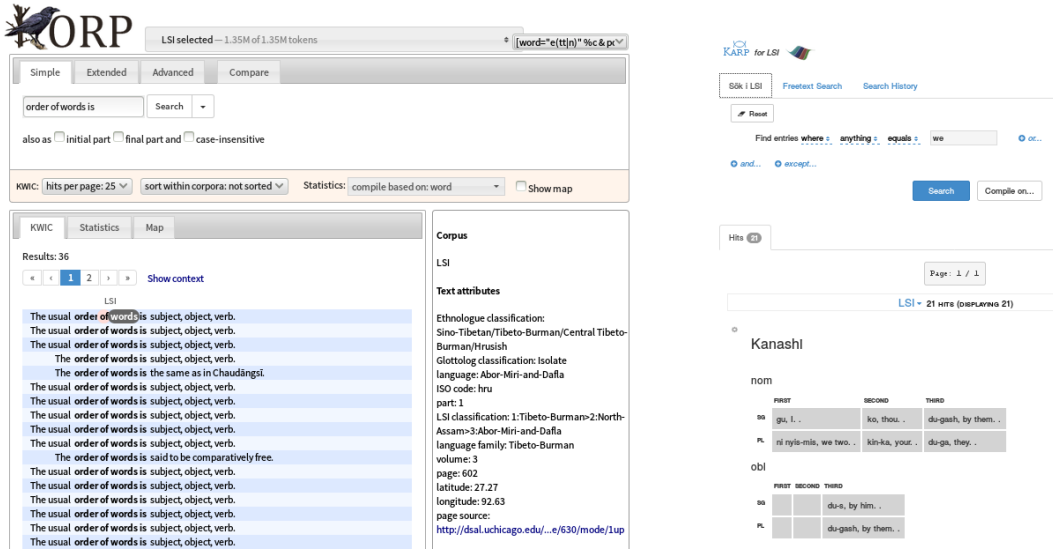

Figure 1: Left: Korp KWIC view • Right: Karp pronoun search

The linguists in the project team will be greatly helped by having access to tools allowing them to browse and search the vast LSI material effectively. This is true for those designing the questionnaires, but in particular and to a much higher degree for those charged with filling out the questionnaires – typically linguistics master students – using the LSI grammar sketches.

For effective exploration of the digitized LSI already in the early stages of this project, and also in order not to spend too much project resources on useful but peripheral tool development, we have strived to reuse existing language tools and infrastructure to the greatest extent possible, even if these tools were not designed explicitly for the kind of large-scale comparative linguistic investigations which are being planned in this project, but rather for more traditional corpus-linguistic studies. Thus, the project team decided to recycle some existing e-infrastructure components, rather than attempting to build a new system from scratch. In the following we describe how this was done.

2 LSI grammar sketches as corpus

After some initial data preprocessing and metadata preparation, the text portions of the descriptive grammars were imported into the Korp corpus infrastructure developed and maintained by Språkbanken (the Swedish Language Bank) at the University of Gothenburg, Sweden. Korp is a versatile open-source corpus infrastructure (Borin et al., 2012b).[1] which is used by several CLARIN centers in the Nordic countries, and also, e.g., in Estonia.

Korp is a modular system with three main components: a (server-side) back-end, a (web-interface) front-end, and a configurable corpus import and export pipeline. The back-end offers a number of search functions and corpus statistics through a REST web service API.

The front-end provides various options to search at simple, extended, and advanced levels in addition to providing a comparison facility between different search results. Search expressions combine text content and text metadata conditions in a unitary search language, and Korp provides deep linking, allowing searches to be bookmarked and shared/reused.

The corpus pipeline is a major component and can be used to import, annotate, and export the corpus to other formats. For annotations, it relies heavily on external annotation tools such as segmenters, POS taggers, and parsers. For our purposes, we have incorporated the English Stanford Parser (Manning et al., 2014) for lexical and syntactical annotations. We have added word and text level annotations to the LSI data, as follows:

Word-level: lemma, part of speech (POS), named entities, normalized word-form, dependency relation.
Text-level (metadata): LSI volume/part, language family, language name, ISO 639-3 language code, longitude, latitude, LSI classification, Ethnologue classification (Lewis et al., 2016), Glottolog classification,[2] page number, page source URL.

Currently, the LSI "corpus" comprises about 1.3 MW, and contains data for around 550 linguistic varieties that we identified during the pre-processing step. Figure 1 (left-hand view) shows a screenshot of the Korp front-end displaying results of a simple corpus query in Korp's KWIC view. The box to the right of the KWIC sentences

[1] https://spraakbanken.gu.se/swe/forskning/infrastruktur/korp/distribution
https://github.com/spraakbanken/korp-frontend/
[2] http://glottolog.org

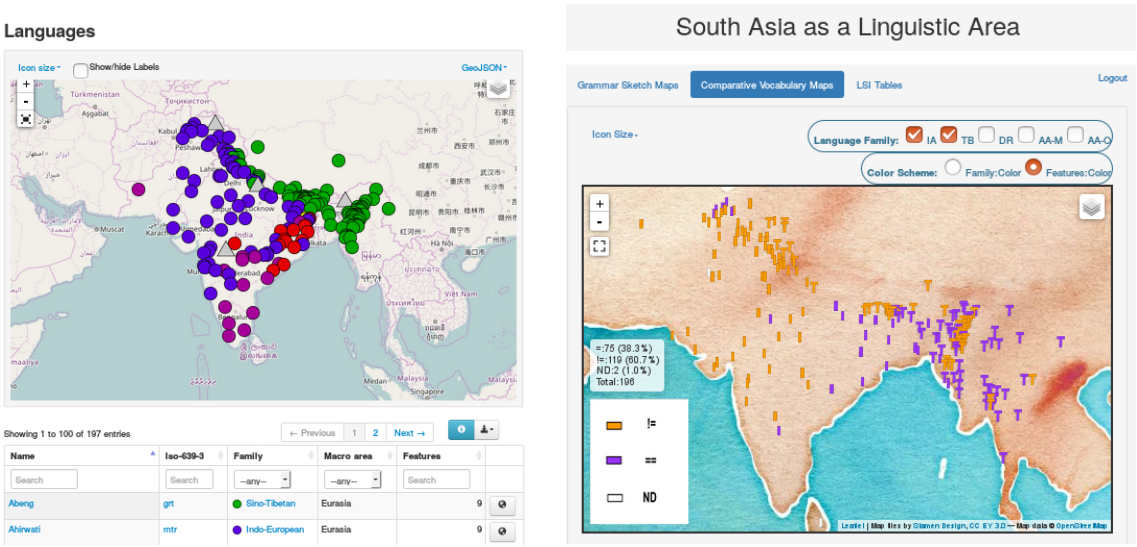

Figure 2: Map visualizations of LSI languages (left) and linguistic features (right)

shows annotations and metadata for the selected word (*Word* and *Text* level attributes), and also provides a link to the corresponding page image available at the *Digital South Asia Library* at the University of Chicago.[3]

The Korp software could be directly used in the project, without any other modification than setting up its configuration files for handling the LSI texts and using English language tools.

3 LSI tables and specimens as lexicons

The LSI grammar sketches contain large amounts of tabular material e.g., inflection tables, pronominal systems, etc., and also language specimens in the form of *interlinear glossed text*, both of which are not particularly suitable for displaying in a corpus KWIC view. Instead, these are imported and stored in another of Språkbanken's infrastructure components, Karp (Borin et al., 2012a). Links are provided from the Korp KWIC metadata box to tables and specimens in Karp, but these can also be accessed directly through the Karp search interface. See Figure 1 (right-hand view).

In the case of Karp, Språkbanken's Karp development team had to be involved in devising an "LSI mode". This fitted well with their ongoing effort aimed at turning Karp – originally designed for browsing and searching Språkbanken's lexical resources – into a more general infrastructure for working with formally structured linguistic data.

4 Visual exploration of the LSI

An important aspect of the linguistic research driving our project is the relationship between *linguistic genealogy* (language family membership), *geography*, and *linguistic features*. Again, the digitized LSI offers such an abundance of data of various kinds, that we need very good tools for exploring this resource for the kind of large-scale comparative linguistic research necessitated by our project objectives. In general, there are indications that data visualization and visual analytics have a crucial role to play in this connection (e.g., Havre et al., 2000; Chuang et al., 2012; Krstajić et al., 2012; Sun et al., 2013).

For the general case, we have adopted the *Cross-Linguistic Linked Data* (CLLD) framework developed by the Max Planck Society,[4] which is open-source and which we could simply install out of the box and configure to display all LSI varieties to which we could assign an ISO 639-3 language code. See Figure 2 (left-hand view).

For the more specific purposes of working with the full LSI data, we have modified the mapping solution available in Korp into an interactive standalone application where the users can view the distribution of linguistic features in LSI varieties on a map. We provide switchable shape/color combinations for visualizing and differentiating family/feature characteristics. Figure 2 (right-hand view) shows a snapshot visualizing the feature **s3sg** ("Is the form of the pronominal 3sg subject the same in intransitive and transitive clauses?", i.e., an indicator of nominative–accusative vs. absolutive–ergative alignment) in languages belonging to the Indo-Aryan and Tibeto-Burman families. The user can select multiple families and multiple features at the same time by checking the

[3] `http://dsal.uchicago.edu/books/lsi/` – page images only; no text search facility is available.
[4] `http://clld.org/`

appropriate check-boxes, and can also switch between color/symbol to visualize feature/family by selecting the appropriate radio button. We have selected feature values to be encoded by color, while the shape of the markers indicate language family (**I** for Indo-Aryan and **T** for Tibeto-Burman. This map indicates that there is a clear areal distribution of this feature in South Asia.

5 Conclusions and future work

Turning the LSI into a structured digital resource will provide a rich empirical foundation for large-scale comparative studies in South Asia. The project described above aims to do this by a combination of manual and automatic information extraction. In order to get the project off the ground quickly, we needed tools for browsing, searching and visualizing the abundance of information present in the LSI. Recycling existing infrastructure components has turned out to be surprisingly effective. We have been able to use Korp and the CLLD framework more or less off the shelf. Rendering the LSI tabular data in Karp required modifications to the Karp infrastructure, and the geographical mapping solution shown in Figure 2 (right-hand view) in practice is a new component developed in this project.

The linguists working with the questionnaires have expressed their satisfaction with Korp as an "information retrieval" interface to the LSI text. An added value in this context is that they have been asked to save the search results – sentences in the text – found by them to be the most relevant to determining a particular linguistic feature, thus providing invaluable input to our work on designing an IE system targeting linguistic information expressed in conventional descriptive grammars.

The status of the project is that most of the LSI has been digitized, the browsing, search and visualization applications described above have been implemented,[5] and the manual questionnaire work and the development of the IE application is underway.

Acknowledgments

The work presented here was funded by the Swedish Research Council as part of the project *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (2015–2019, contract no. 421-2014-969), as well as by the University of Gothenburg and the Swedish Research Council through their funding of the Språkbanken and Swe-Clarin research infrastructures, respectively.

References

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012a. The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 3598–3602, Istanbul. ELRA.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012b. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*.

Karen Ebert. 2006. South Asia as a linguistic area. In Keith Brown, editor, *Encyclopedia of languages and linguistics*. Elsevier, Oxford, 2nd edition.

George A. Grierson. 1903–1927. *A Linguistic Survey of India*, volume I–XI. Government of India, Central Publication Branch, Calcutta.

Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, Salt Lake City.

Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A. Keim. 2012. Incremental visual text analytics of news story development. In *Proceedings of VDA 2012*.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. 2016. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 19th edition. Online version: `http://www.ethnologue.com`.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014*, pages 55–60.

Colin P. Masica. 1976. *Defining a linguistic area: South Asia*. Chicago University Press, Chicago.

Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.

[5] The LSI goes out of copyright towards the end of the project and our data will subsequently be made openly available.