Exploring Big-Data Methods for Large-Scale Comparative Linguistic Research

Lars Borin[1], **Shafqat Mumtaz Virk**[1], **Anju Saxena**[2]

[1]Språkbanken, University of Gothenburg; [2]Linguistics and Philology, Uppsala University
lars.borin|shafqat.virk@svenska.gu.se, anju.saxena@lingfil.uu.se

Abstract

We present ongoing work aiming at turning the linguistic material available in Grierson's 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 language typology and contact linguistics. In this work, we develop state-of-the-art language technology for automatically extracting the relevant information from the text of the LSI.

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. In this work, we develop state-of-the-art language technology for automatically extracting the relevant information from the text of the LSI.

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 volumes of around 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 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.

In addition to the grammar sketches, the core vocabulary also provides some phrases and clauses (e.g., 'good man' $\sim$ 'good woman' $\sim$ 'good men' $\sim$ 'good women', and 'I, thou, etc. go' $\sim$ 'I, thou, etc. went'), making it useful for comparative studies of some grammatical features, in addition to studies of lexical phenomena.

The sheer volume of the LSI material, together with the complexity of processing necessary to extract grammatical information from it and turn the information into a formally structured format useful for large-scale linguistic investigation, warrants the use of the label "big-data methodology" for our endeavor.

2. Data Preparation

2.1 Preprocessing

As a first step, we are in the process of digitizing all LSI volumes dealing with the four main South Asian language families (16 out of the 19 books). This part of the work is almost completed. Since OCR software deals poorly with the complex typography and multitude of languages of the language examples and language specimens in the LSI, the digitization is accomplished by an initial scanning and OCR step, followed by a manual correction step, so-called double keying, done by a commercial provider. During the latter, we deliberately chose not to represent the many diacritic characters appearing in the text in their original shape, but rather replace them with unique character combinations easily entered using an ordinary QWERTY keyboard. However, we want these characters restored back to their original shapes in the text that we will be working with. Also, there was a lot of metadata present on each page, in the form of page headers and footers, that we wanted to separate from the language descriptions. So a natural first step was to do some cleaning and pre-processing. Using a set of regular expressions, and mostly relying on a search and replace strategy, both of the above given tasks were completed. Though the process overall went smoothly, there are still some known issues, having to do with rendering superscript characters and characters with complex combinations of diacritics.

2.2 Text Processing and Annotation

The amount of text that has been digitized so far is well in excess of one million words, and in order to be able to

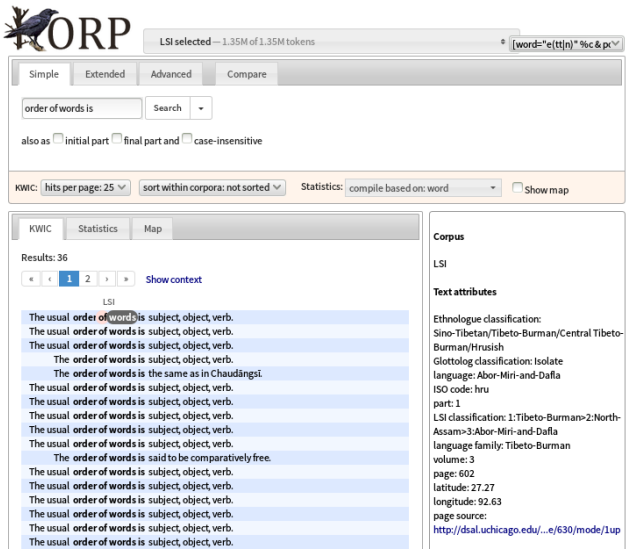

Figure 1: Korp KWIC view resulting from searching the LSI for the string "`order of words is`"

explore this amount of data – which is not feasible to do manually – from the early stages of this project, we have strived to use existing language tools to the greatest extent possible, even if these tools were not designed explicitly for the kind of large-scale comparative linguistic investigations that we have in mind, but rather for more traditional corpus-linguistic studies.

The text data, i.e., grammar sketches excluding tabular data (e.g., inflection tables) and text specimens, have been imported and made searchable using Korp, a versatile open-source corpus infrastructure (Borin et al., 2012).[1] Currently, the LSI "corpus" comprises about 1.3 MW, and contains data about around 550 linguistic varieties that we identified during the pre-processing step. The comparative dictionary and the tabular data from the grammar sketches still remain to be processed in a similar way.

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.

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. Previously, it has mostly been used for Swedish text, and comes with very limited support for English in the vanilla distribution. 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. The following is a list of all annotations that were added:

Word-level annotations: lemma, part of speech (POS), named-entity information, normalized word-form, dependency relation. These are all added automatically.

Text-level annotations: LSI volume/part number, language family, language name, ISO 639-3 language code, longitude, latitude, LSI classification, Ethnologue classification (**?**), Glottolog classification,[2] page number, page source URL, paragraph and sentence level segmentation. These have been added in a semi-automatic manner.

Figure 1 shows a screenshot of the Korp front-end displaying results of a simple corpus query in Korp's KWIC (Key Word In Context) view. The box to the right of the KWIC sentences shows annotations and metadata for the selected word (*Word* and *Text* level attributes).

3. Grammatical Feature Extraction

After having cleaned the LSI data and stored it in a structured way, the next step is to extract information about particular grammatical features of LSI languages. The extracted grammatical feature values are to be used to investigate genetic relations between, and areal influences of languages on each other during the later states of the project. However, in the present paper our focus will be more on methodological development than on linguistic analysis of the results.

We have identified an initial list of features that we think are interesting and will be useful in investigating genealogical and areal influences. For the purpose of extracting values and/or descriptions of those identified features from the LSI data, we have experimented with two different approaches: (1) Pattern based information extraction (IE); (2) Semantic parsing based information extraction. The following two subsections briefly describe each of these two strategies.

3.1 Pattern Based IE

Pattern based IE consists of the following two steps:

(1) Using the standard search API of Korp, retrieve a set of potential sentences from the language descriptions by searching for a particular text string (representing a particular feature) and by limiting the search to 'within sentence'. The extracted sentences are further processed as described below to extract the feature values.

(2) A set of patterns was designed to extract as precisely as possible the feature values. A pattern basically is a regular expression that is used to match and extract the corresponding text segment representing the particular feature values from within the sentences extracted in step 1.

For example, suppose that we are interested to extract information about the normal word order of a particular LSI language from the language description. With step 1, we can extract all sentences having the string "order of words is" from the description of a language (see Figure 1). Next, using the pattern `(.*) (order of words is) (.*)`, one can

[1] `http://spraakbanken.gu.se/swe/forskning/infrastruktur/korp/distribution`

[2] `http://glottolog.org`

first split each sentence into three parts: the part appearing before the string "order of words is", the string itself, and the part appearing after this string. The resulting parts are processed further with more specific patterns (e.g. `(\w+), (\w+), (\w+)`) to extract the 'order of words' of a particular language.

This simple approach allowed us to get off the ground quickly, but it has serious limitations. This pattern based strategy will very strictly match particular sentence structures and/or contents. This probably will not cover all possible ways the same information could have been encoded unless one designs patterns rich enough to catch all possible instances. For such reasons, we have started experimenting with approaches inspired by semantic analysis and *Open Information Extraction* (e.g., Fader et al. 2011).

3.2 IE Based on Semantic Parsing

Shallow semantic analysis or semantic role labelling (SRL) is a process of identification and labelling of semantic roles (also known as semantic arguments) associated with a verbal predicate, a nominal predicate, or a semantic frame depending on the theoretical framework on which an SRL system is based. In our work, we are using SRL to extract values of particular grammatical features from grammar descriptions – a type of information extraction. The procedure is outlined below.

The description of a particular linguistic variety was parsed using a dual layer semantic parser (Ku et al., 2015). From each parsed sentence a list of predicates and their semantic arguments were extracted. The predicates and their arguments were then further processed to find if a particular predicate and its semantic arguments contain the information we are interested in (i.e., information on particular linguistic features and their values). The further processing involved linking of particular predicates to particular features, inspecting the semantic arguments contents, and formulating the feature values and/or description pairs. Suppose for example that we are interested in extracting information about adjective-noun order for a particular LSI language, e.g., Siyin (a Tibeto-Burman language of Burma with about 10,000 speakers). In the LSI description of this language, the information about adjective-noun order has been encoded through the sentence 'The adjectives follow the noun they qualify'. The semantic parse of this sentence will give us the following result:

```
Predicate=follow
ARG0=The_adjectives
ARG1=the_noun
```

The predicate 'follow', in this case, is linked to the noun-adjective order (The other potential candidate predicates for noun-adjective order are 'precede', 'come', etc.). The next step is to examine the semantic arguments of the predicate 'follow', and formulate the feature value. Since, ARG0[3] contains 'adjective(s)', and ARG1 is 'the_noun(s)', we can formulate and return the feature value 'NA' (representing the fact that adjectives follow the nouns). Had ARG0 been 'nouns', ARG1 been 'adjective' with predicate being 'follow', or ARG0 being 'adjectives', ARG1 being 'nouns', and predicate being 'precede', we would have returned 'AN' as the feature value. We have used simple if-else condition based rules to examine predicates and their semantic arguments for the purpose of formulating the feature values. Table 1 shows some statistics about for how many languages and what features values we were able to extract using this strategy. Note that at this stage of our experiments, we have no way of estimating the accuracy of the extraction. We are currently in the process of manually preparing evaluation data for a small number of LSI features.

[3]Traditionally, the semantic arguments of verbs are numbered from ARG0 to ARG5 in addition to the functional modifier arguments such as ARGM-MNR, ARGM-LOC etc. See Palmer et al. (2005) for more details

4. Visualization for Linguistic Research

The work presented here is part of a larger effort to design and deploy language-technology based e-science tools for research in large-scale comparative linguistics, where one important data source consists of massive amounts of text (the LSI grammar sketches and text specimens in our case). There are indications that data visualization and visual analytics will play a crucial role in this connection (e.g., Havre et al. 2000; Smith 2002; Schilit and Kolak 2008; Chuang et al. 2012; Broadwell and Tangherlini 2012; **?**; Sun et al. 2013).

Consequently, one of the objectives of this project is to develop state-of-the-art tools for visualization of the extracted grammatical features in a way that makes it easy to observe the genetic relations between multiple languages and the areal influences of languages on each other. For this purpose, we have developed an interactive mapping solution where the users can choose to view values of particular feature(s). Further, we provide switchable shape/color combinations for visualizing and differentiating family/feature characteristics. Figure 2 shows a snapshot visualizing the feature nlpos (i.e. numeral position w.r.t noun) for the languages belonging to the Tibeto-Burman family. As can be noticed, we have selected feature values to be encoded by the color, while the symbol 'T' shows that these languages belong to the Tibeto-Burman family. The user can select multiple families and multiple features at the same time by checking the appropriate check-boxes, and can also switch between color/symbol to visualize feature/family by selecting the appropriate radio button. From this map, it can be observed that the numerals mostly follow the noun (indicated by brown color) in those languages spoken on the southern areas, while it mostly precedes (indicated by purple color) in the languages spoken on the middle and northern parts. Such interactive mapping facility provides a useful way to show the genetic relations and areal influences between languages spoken in different geographical areas and belonging to different language families.

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. Further, such studies cannot be conducted manually, but need to draw on extensive digitized language resources and state-of-the-art computational tools. This is the main goal of our on-going work with the

Feature	Value (# of languages)	Value (# of languages)	Value (# of languages)	Total
Adj pos.	A-N (19)	N-A (42)	BOTH (5)	66
def. article	YES (21)	NO (7)		28
indef. article	YES (72)	NO (0)		72
honorific	Verb (1)	Pronoun (8)	Other (34)	43
number base	10 (4)	20 (17)	both (7)	28
numeral pos.	#-N (9)	N-# (25)		34
N number	TWO-WAY (41)	THREE-WAY (19)		60
reflexive	verb (2)	pronoun-1 (31)	pronoun-2 (3)	36

Table 1: Statistics of Retrieved Features

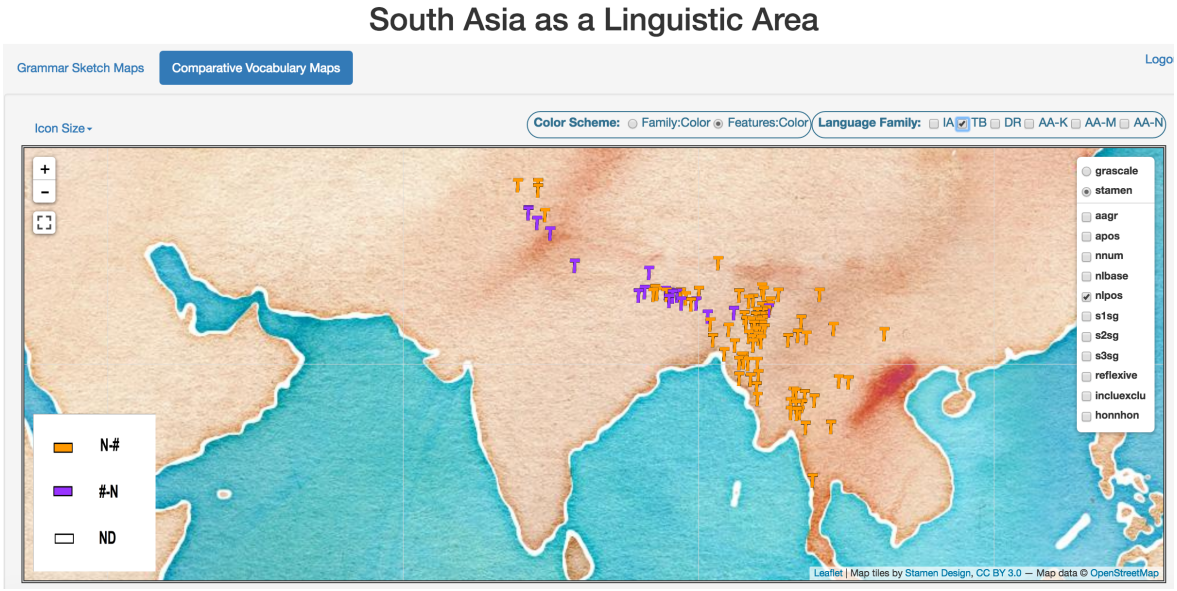

Figure 2: Map showing numeral position w.r.t noun

LSI. In addition to this, we aim to contribute to the methodological development of large-scale comparative linguistics drawing on digital language resources, as well as to the methodological development of SRL based and open information extraction, adapting these paradigms to a different and hitherto unexplored domain. In the longer perspective, we hope that the solutions which we develop in our work will be more generally applicable to the text mining of descriptive grammars – which are increasingly available in digital form – so that the resulting formally structured linguistic information can be used to populate linguistic databases.

In the future, we would also like to take into account the phonological and other related information present in tabular data and the parallel annotated data present in the text specimens provided with LSI grammar sketches.

References

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

Peter M. Broadwell and Timothy R. Tangherlini. 2012. TrollFinder: Geo-semantic exploration of a very large corpus of Danish folklore. In *The Third Workshop on Computational Models of Narrative*, pages 50–57, 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.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP 2011*, pages 1535–1545, Edinburgh. ACL.

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.

Lun-Wei Ku, Shafqat Mumtaz Virk, and Yann-Huei Lee. 2015. A dual-layer semantic role labeling system. In *ACL*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March.

Bill N. Schilit and Okan Kolak. 2008. Exploring a digital library through key ideas. In *Proceedings of JCDL'08*, pages 177–186, Pittsburgh. ACM.

David A. Smith. 2002. Detecting and browsing events in unstructured text. In *SIGIR'02*, Tampere. ACM.

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.