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 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 to language typology and contact linguistics. The project has two concrete 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 provided by Swe-Clarin, the Swedish CLARIN consortium, 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 tools for information extraction from the descriptive grammars.

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

1.1 South Asian Linguistics and the Areal Hypothesis

South Asia (also "India[n subcontinent]") with its rich and diverse language ecology and a long history of intensive language contact provides abundant empirical data for studies of linguistic genealogy, linguistic typology, and language contact.

This region (normally understood in linguistic works as comprising the seven countries Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka, as well as adjacent areas in neighboring countries, since language boundaries do not always coincide with national borders) is the home of hundreds of languages spoken by almost two billion people – more than a quarter of the world's population. Most of the 661 living languages of South Asia (Simons and Fennig, 2018) are from four major language families (Indo-European>Indo-Aryan and Nuristani, Dravidian, Austroasiatic>Munda, Khasian and Nicobaric, and Tibeto-Burman (Sino-Tibetan); see Figure 1). In addition there are some language isolates and small families (Georg, 2017) and several creoles and pidgins.

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 only (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* into a digital language resource, a database suitable for a broad array of linguistic investigations of the languages of South Asia, and especially for conducting a more thorough scrutiny of the South Asian areal hypothesis.

Given the CLARIN context, we will focus on some research infrastructural aspects of our work here, notably how the project was able to reap great benefits from repurposing existing infrastructure components provided by Swe-Clarin, the Swedish CLARIN consortium.

Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017.
Conference Proceedings published by Linköping University Electronic Press
at www.ep.liu.se/ecp/contents.asp?issue=147. © The Author(s).

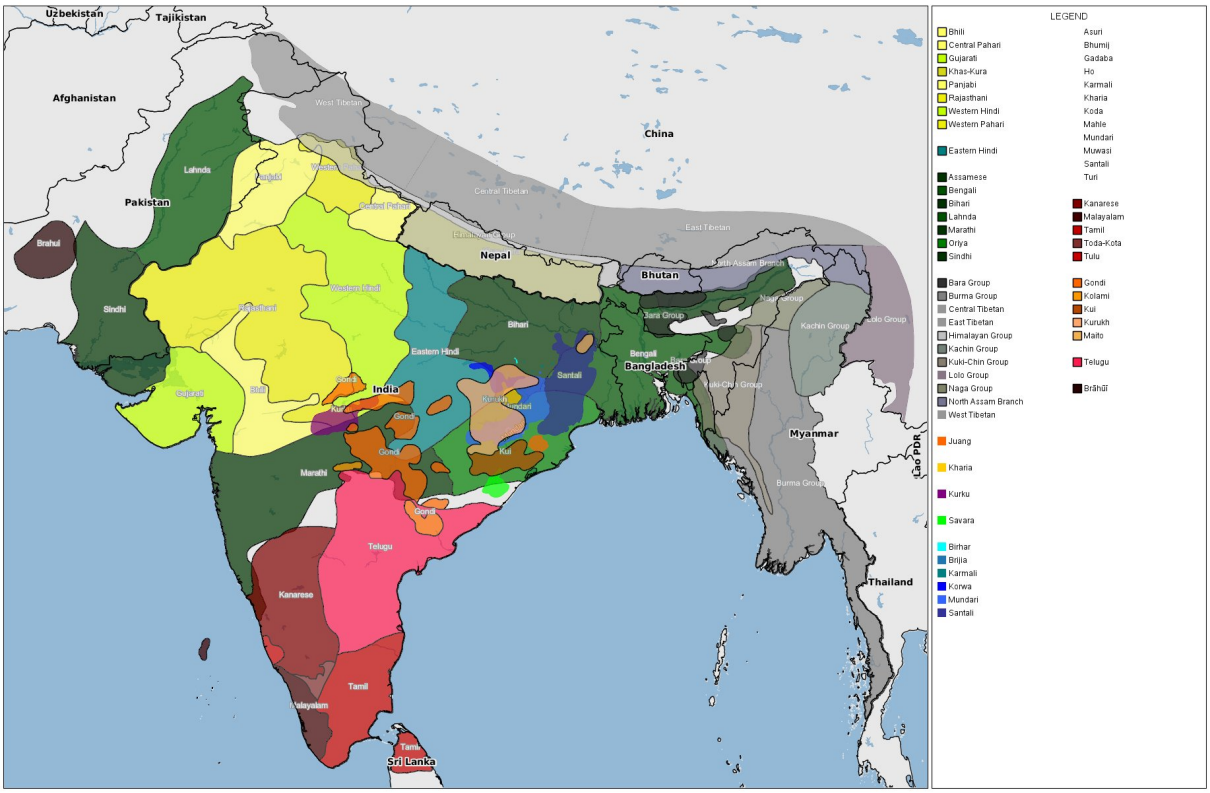

Figure 1: The four major language families of South Asia (from `http://llmap.org`)

1.2 Grierson's *Linguistic Survey of India*

The linguistic richness and diversity of South Asia was documented by the British government in a large-scale survey conducted in the late nineteenth and the early twentieth century under the supervision of Sir George Abraham Grierson and Sten Konow. The survey resulted into a detailed report comprising 19 volumes of around 9,500 pages in total, entitled *Linguistic Survey of India* (LSI; Grierson, 1903–1927). The survey covered 723 linguistic varieties representing the major language families of the region 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 grammar 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 LSI grammar sketches provide basic grammatical information about the languages in a fairly standardized format. The focus is on the sound system and the morphology (nominal number and case inflection, verbal tense, aspect, and argument indexing inflection, etc.), but as we will see below in Section 3, there is also some syntactic information to be found in them. Importantly, the 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, converbs/conjunctive participles, but go considerably beyond these, offering the possibility of a broad comparative study of South Asian languages.

The grammar sketches range in length from less than a page to over eighty pages, and the whole LSI comprises far too much text for it to be a realistic option to process it manually. Arguably, this language data qualifies as "big data", not primarily by virtue of its volume, but by virtue of its inherent complexity and the tools needed to process it (Ward and Barker, 2013), especially for extracting and comparing grammatical features.

Thus, we are currently exploring information extraction methodologies which could help us turning the free-text descriptions of the LSI grammar sketches into formally structured tabular data suitable for large-scale automatic processing. At the present time, this is the main NLP focus of the project. Since

the grammatical descriptions are written in English, this of course means that the information extraction application that we are developing will be for English (see Section 3 below).

The core word lists which accompany the language descriptions are collected in a separate volume (Volume 1, Part 2: *Comparative vocabulary*). Each list holds a total of 168 entries. Most of the entries in the comparative vocabulary render concepts which cover a broad spectrum consisting of body parts, domestic animals, personal pronouns, numerals, and astronomical objects. There is some overlap with other concept lists used in language classification: For instance, 38 of the concepts are also found in the shorter (100-item) version of the so-called *Swadesh lists*, core vocabulary lists originally devised by the American linguist Morris Swadesh (1955) specifically for the purpose of inferring genealogical relationships among languages. Thus, the LSI comparative vocabulary clearly has one part that can be used in investigating genetic connections among the languages, but also another part – at least half of the entries – which we hypothesize could be used to find areal influences.

Notably, the LSI comparative vocabulary also provides some phrases and propositions (e.g., 'good man' ∼ 'good woman' ∼ 'good men' ∼ 'good women', and 'I, thou, etc. go' ∼ 'I, thou, etc. went'), making it useful for comparative studies of some grammatical features, in addition to studies of lexical phenomena. In a preliminary study, some grammatical features have been semiautomatically extracted from the comparative vocabulary, and used as a kind of "silver standard" in some of our information extraction experiments.

The language data for the LSI grammar sketches were collected around the turn of the 20th century, hence obviously reflecting the state of these languages of about a century ago. However, we know that many grammatical characteristics of a language are quite resistant to change (Nichols, 2003), much more so than vocabulary. In order to get an understanding of the usefulness of the LSI for our purposes, we sampled information from a few of the sketches in order to see how well the LSI data reflect modern language usage. Our results show that while some of the lexical items are not used today in everyday speech, most other information reflects in many ways the modern language, and thus cannot be treated as representing an 'archaic' variety of, e.g., Hindi.

Despite its age, LSI still remains the most complete single source on South Asian languages. It has been used in a few studies with varying aims and objectives, but has not been exploited to the extent it could have been, important reasons arguably being its vast size and limited accessibility. This multi-volume work will generally be found only in select research libraries and it does not have any kind of index of its contents. A scanned version of LSI is now available on the University of Chicago's *Digital South Asia Library* website,[1] although the page images displayed there are neither searchable nor digitally processable, effectively making this version equivalent to the printed LSI w.r.t. accessing its contents, although of course universally accessible to anybody with an internet connection. One of the major objectives of the study reported in this paper is to convert LSI into a digital resource stored in a way which makes it easy to access, explore, and process for deeper linguistic investigations of the languages described in LSI. This digital resource will have rich formally structured metadata as well as the full original text of the LSI.

1.3 Project Aims

On the linguistic side, the major objective of the project is to investigate the claim about South Asia as a linguistic area. The examination of genealogical, typological and areal relationships among South Asian languages requires a large-scale comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools. As mentioned already, there have been some earlier attempts to use LSI in areal studies (e.g., Hook, 1977), but because of the manual nature of these studies, the information in the LSI was used only to a very limited extent, and the results presented in a general, non-concrete manner. Further, no accompanying methodological discussion was offered (e.g., how the data was extracted and analyzed, and for which languages, etc.). We aim to investigate the *South Asia as a linguistic area* claim on the basis of a much broader array of linguistic data using state-of-the-art

[1] `http://dsal.uchicago.edu/books/lsi/`

computational techniques and tools in this study. However, in this paper, we focus on the automatic extraction of linguistic features from the LSI data and the development of a typological database which can be used as a major source for the investigation of the above mentioned claim later (Section 3).

The development of general purpose methodologies and tools for large scale comparative linguistics and visualization of linguistic information is another primary aim of the project (Section 4). Our hope and aim is to build methodologies and tools which will be applicable not only to the LSI grammar sketches, but also to the multitude of descriptive grammars of the world's languages that are digitally accessible and available for linguistic investigations (see `http://glottolog.org`).

The full text of the LSI (only the Latin-script portions) has been 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.

This will be the first large-scale digital resource on South Asian languages which will be completely automated, with a solid 'deep' structure with the possibility of doing searches for grammatical (morphological and syntactic) as well as lexical features, with links to the original LSI pages as well as rich visualizations. Building a database of this magnitude will also contribute at least indirectly to developing NLP tools for South Asian languages. Studies investigating a multitude of linguistic questions relating to lexicon, morphology, syntax, language contact between two specific languages as well as questions relating to areal linguistics and language change will benefit from this resource. We are already using the resource in our linguistic investigations, as we are building the database (cf. Borin et al., 2014; Saxena, 2016).

We also intend to initiate experiments for utilizing the text specimens for extracting additional linguistic data from the LSI, using the English version of the text as pivot, e.g., inferring basic subject–object marking through cross-language annotation projection (see, e.g., Xia and Lewis, 2007).

2 Recycling Research Infrastructure Components

In the first phase of the project, the linguists in the project team have 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 so that they could 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 has been 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.

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 – several of which were available through the Swe-Clarin infrastructure – rather than attempting to build a new system from scratch. In the following we describe how this was done.

2.1 LSI Grammar Sketches as Corpus

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

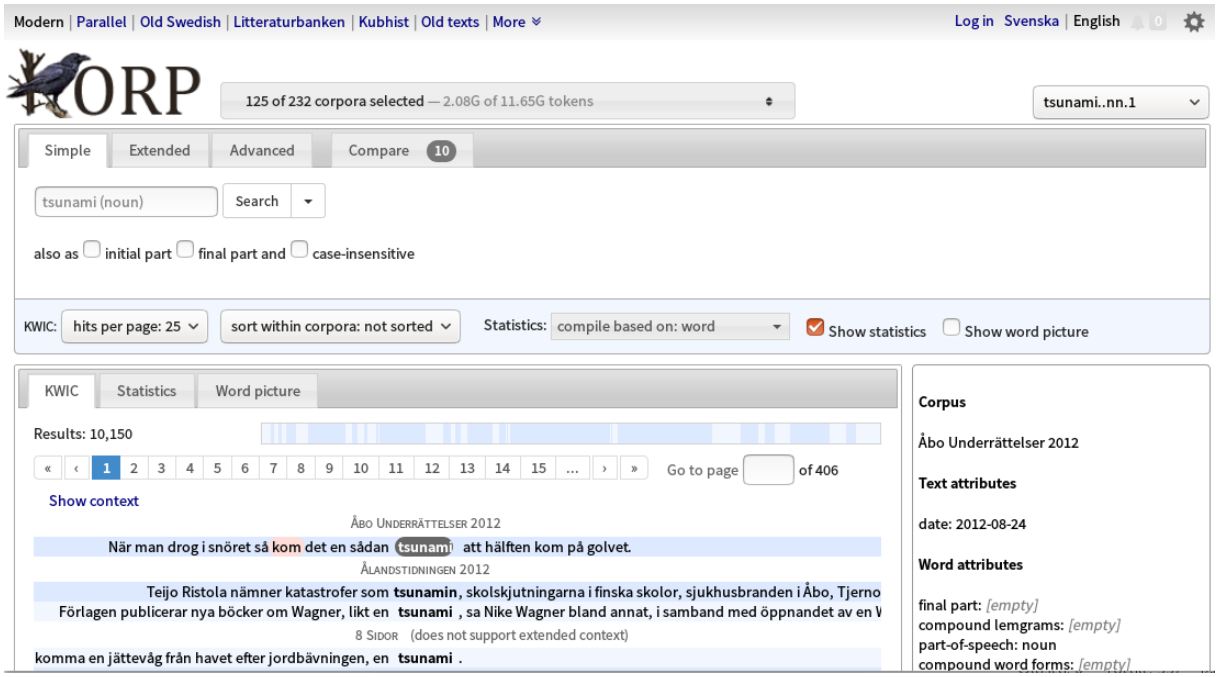

Figure 2: The user interface to Korp, Språkbanken's and Swe-Clarin's corpus infrastructure

et al., 2012b; Hammarstedt et al., 2017a; Hammarstedt et al., 2017b),[2] developed and maintained by Swe-Clarin leading partner Språkbanken (the Swedish Language Bank at the University of Gothenburg). Korp is used by several CLARIN centers in the Nordic countries, and also, e.g., in Estonia. Currently, the LSI corpus comprises about 1.3 million words, and contains data about around 550 linguistic varieties that we identified during the pre-processing step.

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 (Hammarstedt et al., 2017b). The back-end offers a number of search functions and corpus statistics through a REST web service API. As the main corpus search engine, it uses Corpus Workbench (Evert and Hardie, 2011).

The front-end – an in-house development – provides various options to search at simple, extended, and advanced levels in addition to providing a comparison facility between different search results (Hammarstedt et al., 2017a). See Figure 2.

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 pre-existing 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 (Simons and Fennig, 2018), Glottolog classification,[3] page number, page source URL, paragraph and sentence level segmentation. These have been added in a semi-automatic manner.

While most of the annotations are self-explanatory, there are a few which may need some explanation. The *normalized word form* is the form produced by removing the diacritics and other phonological char-

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

[3] http://glottolog.org

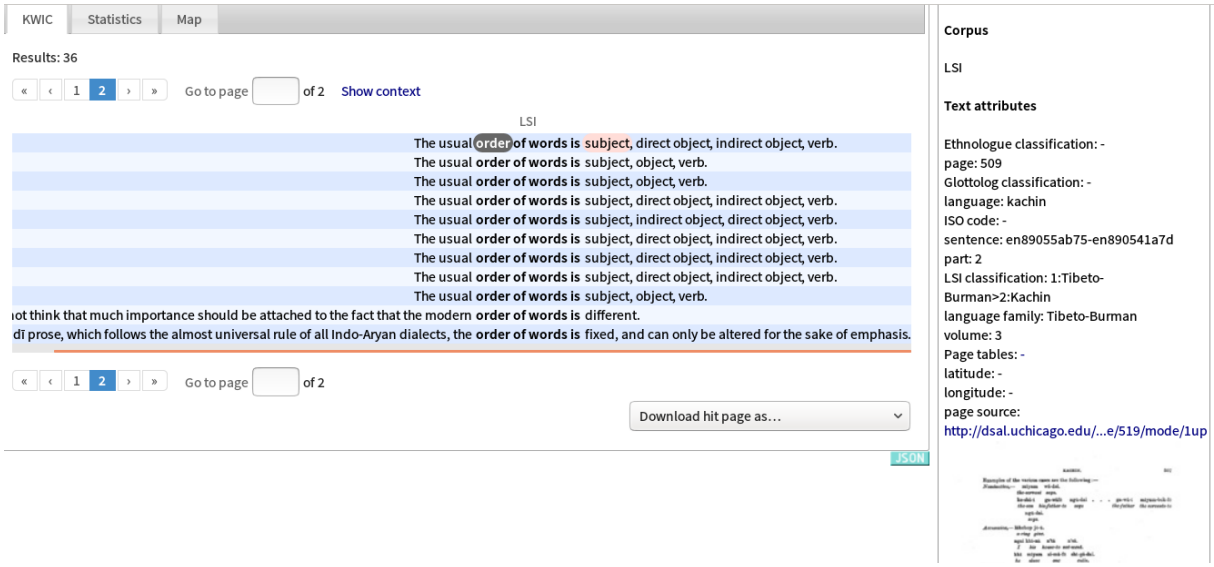

Figure 3: Korp KWIC view resulting from searching the LSI for the string "`order of words is`"

acters. The purpose is to make it easy to search the corpus by using the standard English keyboard without requiring the user to enter accented characters, since the LSI consistently renders language names and glosses in a kind of phonetic transcription which will most likely be unfamiliar to many users. Thus, the normalization allows the user to search for, e.g., *Bihārī* using the search string "Bihari", or "bihari" (with case-sensitivity disabled).

The text-level annotations above mostly represent the metadata which were collected from different sources[4] in addition to the LSI volumes themselves, and are maintained as part of the corpus. The page source URL, for example, is a link to the image version of the corresponding LSI page available from the University of Chicago's *Digital South Asia Library*.

Figure 3 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 query is one aiming at finding out about the basic word order of a language, which is one of the more prominent linguistic features utilized in typological language classification. This illustrates how this basic corpus tool can be repurposed for pursuing the kind of research questions that our linguist project members are interested in. It is true that Korp is not the ideal tool for this, but building a bespoke application would have been far beyond the means of the project.

The box to the right of the KWIC sentences 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.[5]

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.

2.2 LSI Tables and Specimens as Lexicons

The LSI grammar sketches contain large amounts of tabular material, e.g., inflection tables, personal pronoun 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.

Like Korp and other infrastructure components developed and maintained by Språkbanken, Karp, too, is structured with a server backend, a JSON-based web service API and a web application frontend. This means that functionality can be conveniently modularized. For instance, the Korp frontend calls

[4]For instance, location data come mainly from the Glottolog: `http://glottolog.org`.

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

the Karp backend for the lexical information needed in order to execute lemma-based corpus searches, and conversely, the Karp frontend calls the Korp backend in order to offer example sentences for lexical entries.

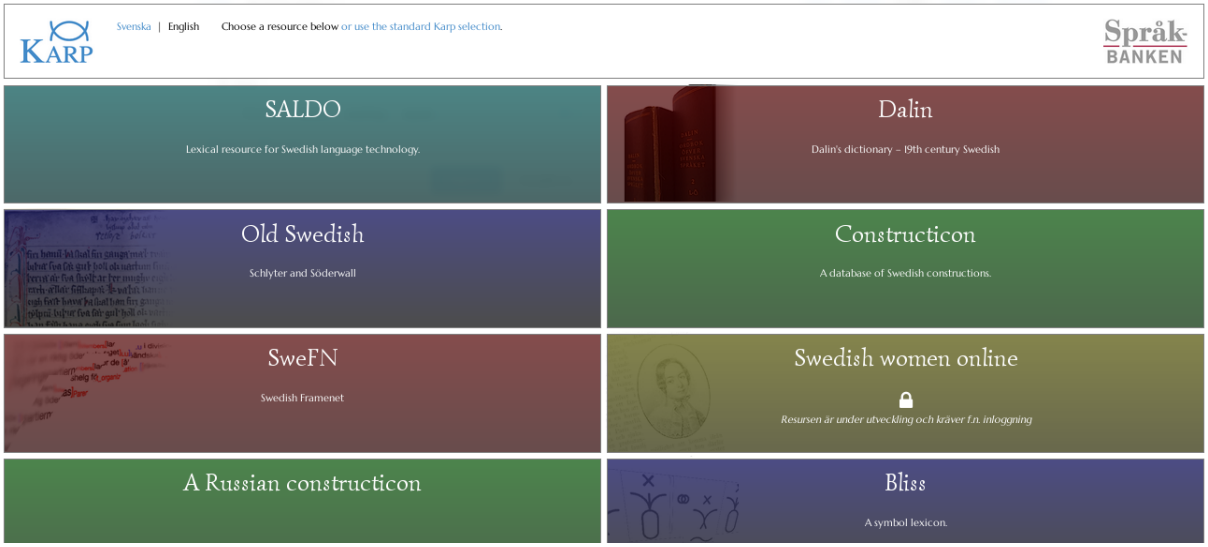

Figure 4: Karp, Språkbanken's and Swe-Clarin's infrastructure for accessing and editing lexical resources and other formally structured language data

Karp started out as a fairly run-of-the-mill web-based search and browsing environment for lexical data. All of Språkbanken's digital lexical resources are available through it, whether born digital – as SALDO, the main lexical resource used for Swedish morphological annotation in the import pipeline of Korp (Borin et al., 2013a) – or digitized versions of traditional dictionaries, including a number of Swedish historical dictionaries (Borin and Forsberg, 2011).

From its humble beginnings, Karp has developed into an infrastructure component for working with language data that has a formally defined (tabular) structure. In addition to lexical entries, this includes also data such as grammatical paradigm tables and encyclopedia articles.

Karp has an editing mode (Borin et al., 2013b), which was originally developed for building the Swedish FrameNet (Borin et al., 2010), but which has since been extended into a general editing environment for formally structured language data. Notably, the Swedish and Russian constructicons are built using the Karp editor (Lyngfelt et al., 2012; Janda et al., to appear 2018), as is the *Swedish Women Online* biographical database.[6] See Figure 4.

In the context of the present project, Språkbanken's Karp development team had to be involved in devising an "LSI mode", but in fact this fitted well with our ongoing effort aimed at turning Karp into a more general infrastructure for working with formally structured linguistic data, as described above. See Figure 5, illustrating a query aiming at at finding out some linguistic features of the personal pronominal systems of the LSI languages. Again, as in the case of Korp mentioned above, this solution is not perfect, but it could have it up and running in a very short time compared to what it had meant to implement the perfect functionality from scratch, meaning that limited project resources can be put to better use, such as deeper linguistic analysis of the LSI data.

3 Automatic Extraction of Linguistic Information from Linguistic Descriptions

After having pre-processed the LSI data and stored it in a structured way, the next step is to extract information about particular grammatical features, and to build a typological database of the LSI languages. The developed feature database is to be used for investigation of the claim about South Asia as a linguistic area during the later stages of the project.

Automatic extraction of linguistic features from traditional linguistic descriptions is a novel task, and has high potential value in typological, genealogical, historical, and other related areas of linguistics

[6] https://skbl.se/

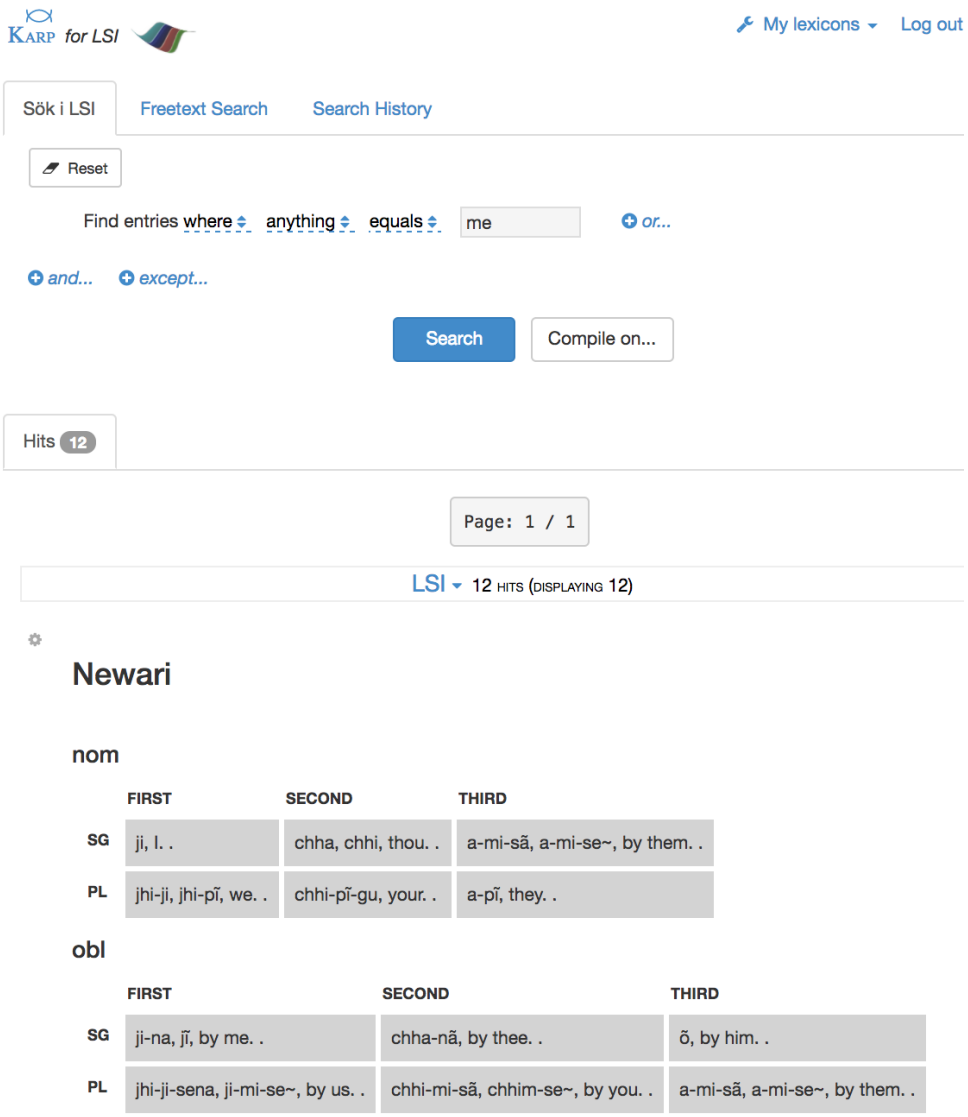

Figure 5: Karp view showing LSI tables of personal pronoun paradigms

that make use of databases of structural features of languages. There exist many typological databases of linguistic structures, including the *World Atlas of Language Structures* (WALS) (`wals.info`), the *Atlas of Pidgin and Creole Language Structures* (APiCS) (`apics.org`), the *South American Indigenous Language Structures* (SAILS) (`sails.clld.org`), AUTOTYP (`github.com/autotyp/autotyp-data`), and the *Phonetics Information Base and Lexicon* (PHOIBLE) (`phoible.org`). To the best of our knowledge, all the linguistic databases published so far have been manually constructed and curated, where human experts have turned information from field data or analyzed data into datapoints in the database. The use of human expertise guarantees a certain level of quality and robustness, but is highly labor intensive and consequently costly. There are some 6,500 languages in the world, out of which descriptive grammars – ranging from brief grammar sketches to multi-volume reference grammars – are available for over 4,000 (see `glottolog.org`). Manually extracting information about 200–300 features from each of them is a very ambitious – and in practice unrealistic – undertaking.[7] Significant amounts of analyzed language data (grammatical descriptions in discursive textual form) are increasingly

[7] Very relevant in this connection is the fact that one of the most ambitious and well-known linguistic-feature datasets, the *WALS* (Dryer and Haspelmath, 2013), even though it reports values for a total of 192 linguistic features in 2,679 languages, in reality most cells in the resulting matrix are empty. In version 2014 of the dataset available for download from `http://wals.info/download`, out of a total of 514,368 cells, no less than 437,903 are empty, meaning that less than 15% of the potential values have actually been filled.

being made available in digital form, and the field of natural language processing (NLP) offers tools that potentially can aid us in extracting information about linguistic features from such textual sources, at least for sources in English and some other languages. To take advantage of these advancements and to help the linguistic community in populating the linguistic feature databases, we have developed methods to automatically extract linguistic features from linguistic grammars.

For our initial study, we have identified a list of features that we think are interesting and will be useful to meet the objectives of the project. Some of these features are:

(1) Apos: What is the order of adnominal property word and noun?[8]
(2) NLpos: What is the order of numeral and noun in the NP?
(3) NLBase: What is the base of the numeral system?
(4) Aagr: Can an adnominal property word agree with the noun in number and/or gender?
(5) AagrNum: Can an adnominal property word agree with the noun in number?
(6) AagrGen: Can an adnominal property word agree with the noun in gender?
(7) Reflexive: What kind of reflexive construction does the language have?
(8) DefArticle: Are there definite or specific articles?
(9) WOrder: What is the order of words?

For the purpose of extracting values and/or descriptions of these features from traditional reference grammars, including the grammar sketches in the LSI, we have experimented with three approaches to information extraction: (1) Pattern based; (2) dependency parsing based; and (3) semantic parsing based. In the following sections, we briefly describe each of the approaches and their results while leaving the details to be reported separately.

3.1 Pattern Based Feature Extraction

The pattern based feature extraction methodology is inspired by pattern based information extraction in general, and predicated on the observation that information about particular linguistic features in descriptive grammars is often given using particular descriptive patterns (at least we have observed this in the case of LSI). Taking advantage of this, one can look for the existence of particular keywords (or a combination of words) within the descriptive grammar to reach to the relevant text, and then process it further to extract the feature value of interest. We used a similar type of two-stage approach to retrieve the relevant sentences from our data using Korp's standard search API first, and then to process them further using regular expression based patterns to extract the feature values. Suppose for example that we are interested in extracting information about the normal word order in a particular LSI language from the language description. As a first step, we can extract all sentences having the string "order of words is" from the description of a language (see Figure 3). Next, using the pattern `(.*) (order of words is) (.*)`, one can 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 can be processed further with more specific patterns (e.g. `(\w+), (\w+), (\w+)`) to extract the 'order of words' of that 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 experimented with approaches inspired by syntactic and semantic analysis, and *Open Information Extraction* based techniques (e.g., Fader et al., 2011).

[8] An *adnominal property word* corresponds to an adjective or participle in English and many other languages.

3.2 Dependency Parsing Based Feature Extraction

Dependency parsing provides syntactic dependency information for the words of a text, which we exploit to extract feature values in this feature extraction strategy. After retrieving the relevant sentences using Korp's search facility (as exemplified above), the sentences were parsed using the Stanford dependency parser (Manning et al., 2014), and the resulting dependencies were further processed using a set of rules to extract the required feature values. Again as an example, suppose that we are interested in extracting information about the order of adjective and noun in the Siyin[9] language. Using Korp's standard search interface, we can extract all sentences containing the lemma "noun" or "adjective" from the language description. One of the extracted sentences will be:

The adjectives follow the noun they qualify .

When we parse this sentence with the Stanford dependency parser, it will return the following dependencies:

```
det(adjectives-2, The-1)
nsubj(follow-3, adjectives-2)
root(ROOT-0, follow-3)
det(noun-5, the-4)
dobj(follow-3, noun-5)
nsubj(qualify-7, they-6)
acl:relcl(noun-5, qualify-7)
```

These dependencies can be processed further with a set of rules to extract the required information. We have worked out a specific set of rules (in the form of an algorithm) for each feature value that we are interested in. The details are beyond the scope of this paper, and will be reported elsewhere.

Table 1 shows how accurately the proposed feature extraction methodology was able to extract different feature values. For each feature, the accuracy value was computed using the following simple formula:

$$Accuracy = \frac{N_{correct}}{N_{extracted}}$$

Where $N_{correct}$ is the number of languages for which the feature value was correctly extracted, and $N_{extracted}$ is the total number of languages for which the feature value was extracted. To decide if an extracted value is correct or not, it was compared to the gold value which was retrieved manually by a human expert from the comparative vocabulary or from the language descriptions.

Feature	Accuracy (%)
Apos	0.818
NLPos	1.0
NLBase	0.823
Reflexive	0.739
Aagr	0.857

Table 1: Evaluation results: Dependency parsing

Once again, this strategy will very strictly match particular sentence structures and contents of arguments. To address some of the limitations of this strategy, we report another strategy in the next subsection which is based on semantic parsing.

3.3 Semantic Parsing Based Feature Extraction

Shallow semantic analysis or semantic role labelling (SRL) is the process of identifying and labeling the semantic roles (also known as semantic arguments) associated with verbal or nominal predicates in a

[9]Siyin (csy) is a Tibeto-Burman language spoken in Burma.

Predicate	Semantic arguments
follow	ARG1:The_adjectives, ARG2:the_noun_they_qualify
qualify	ARG1: the_noun_they

Table 2: Semantic parse

given piece of text. Automatic semantic role labeling finds applications in many areas of NLP including information extraction (Surdeanu et al., 2003), and in this work we are using it for feature extraction – a sort of information extraction. In this strategy, after having parsed the sentences using a semantic parser (Björkelund et al., 2009), the parses are further processed to extract feature values. The further processing steps involve (1) checking for particular predicates for particular features; (2) inspecting the semantic arguments' structure and contents; and (3) formulating the feature values. Using our previous example, i.e., the ordering of adjective and noun in the Siyin language, this time we can semantically parse the sentence *The adjectives follow the noun they qualify* to get the verbal predicates and their semantic arguments as given in Table 2.

The predicate 'follow' is one of those predicates that we had identified, independently, to be linked to the adjective–noun order feature: Using a development data set, we identified a set of predicates linked to each of the target features. This simply involved finding sentences in the descriptive grammars which were used to provide information about a particular feature, and then analyzing them to find the associated list of predicates.

The next step is to examine the semantic arguments of the predicate 'follow', and formulate the feature value. According to Propbank,[10] for the predicate 'follow', ARG1 represents the thing following, while ARG2 represents the thing followed. In the analysis shown in Table 2, the string *The adjectives*, is ARG1 (i.e. the thing following), while the string *the nouns they qualify* is ARG2 (i.e. the thing followed). The substrings representing ARG1, and ARG2 can be further analyzed to formulate and return the feature value '2-N-ANM' (the fact that adjectives follow the nouns). Had ARG1 contained *noun(s)*, ARG2 contained *adjective(s)* with predicate being 'follow', or ARG1 contained *adjective(s)*, ARG2 contained *noun(s)*, and the predicate being 'precede', '1-ANM-N' (the fact that adjectives precede nouns) would have been returned as the feature value. We have used simple if-then-else conditions to examine predicates and their semantic argument strings for the purpose of extracting and formulating the feature values. In the future, we plan to experiment with more advanced techniques, such as active learning, for the feature extraction and formulation from the semantic parses.

In order to test the generality of our approach, for this experiment, rather than drawing on the LSI grammar sketches, we have worked with digitized reference grammars used in the Grambank project,[11] for a set of languages where the linguistic features of interest have already been extracted manually from exactly the grammars used in our experiment, thus providing us with a gold standard dataset.

The evaluation results of this strategy are given in Table 3. As can be seen, the system has varying precision and recall for different features, which highlights the difficulty/ease of automatically extracting the corresponding feature values using the described method.

As mentioned previously, there does not exist any work related to automatic linguistic feature extraction, which means we do not have any other system to compare the proposed system's performance. Instead, we evaluate the system performance against a baseline calculated for each feature on the basis of the most frequent feature value. As can be noted, for four out of the five features, the proposed system was able to easily beat the baseline precision values, the exception being the feature 'AagrNum'.

[10]The lexico-semantic resource on which the semantic parser is based

[11]Grambank is an ongoing initiative at the Max Planck Institute for the Science of Human History at Jena, developing a database of structural (typological) features for a substantial part of the world's languages. The grammars and gold-standard feature sets used in this experiment were kindly put at our disposal by Harald Hammarström.

Feature	Precision	Recall	F-Score	Baseline Precision
Apos	0.76	0.40	0.52	0.41
NLpos	0.85	0.30	0.44	0.75
AagrNum	0.69	0.21	0.32	0.77
AagrGen	0.64	0.14	0.23	0.27
DefArticle	0.84	0.27	0.41	0.27

Table 3: Evaluation results: Semantic parsing

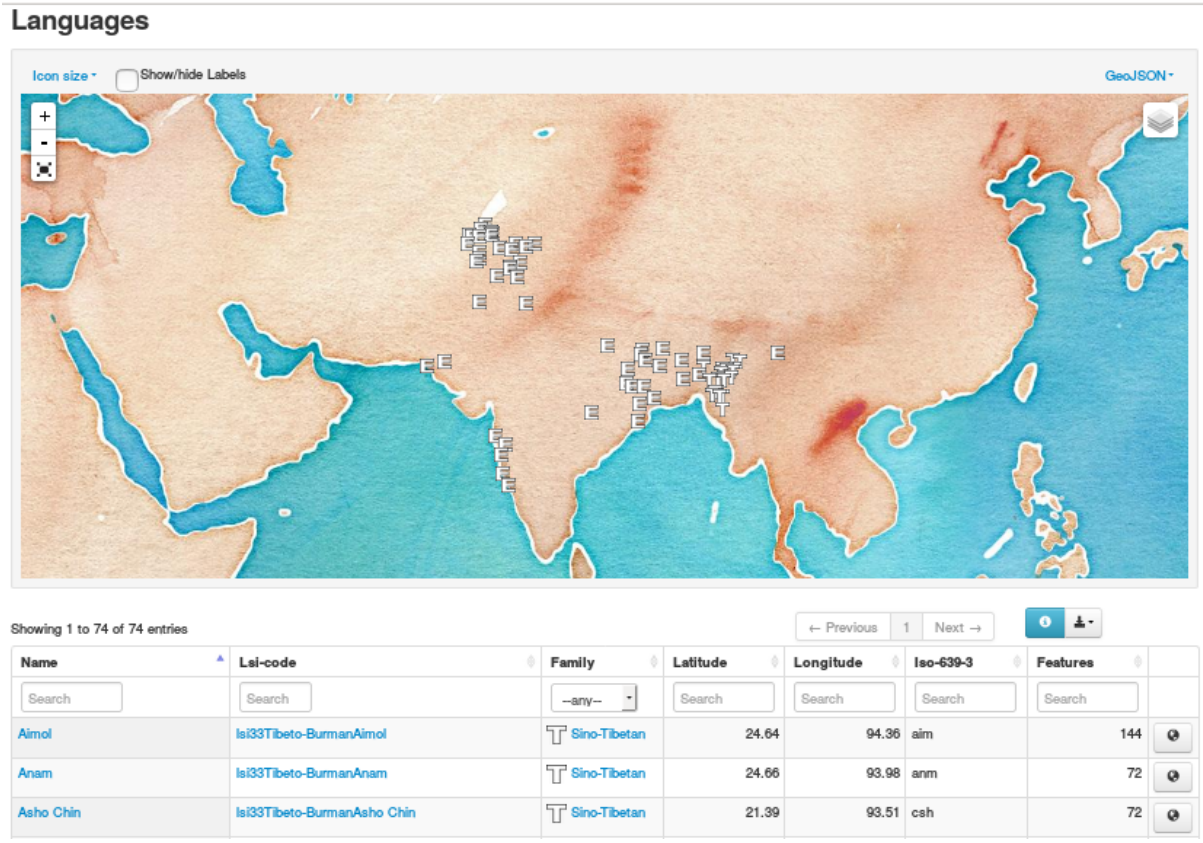

Figure 6: The LSI dataset in CLLD

4 Visualization for Linguistic Research: 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. 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). So, we are developing a number of solutions for better visualization of languages and their features on maps to help the linguistic community working in the areas mentioned above, and to achieve the goals of the LSI project.

For the general case, we have adopted the *Cross-Linguistic Linked Data* (CLLD) framework developed by the Max Planck Society,[12] 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 6.

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

[12] `http://clld.org/`

of linguistic features in LSI varieties on a map. We provide switchable shape/color combinations for visualizing and differentiating family/feature characteristics. Figure 7 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 appropriate check-boxes, and can also switch between color/symbol to visualize feature/family by selecting the appropriate radio button. In the map n Figure 7 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 in Figure 7). In this map we can discern a clear areal distribution of this feature in South Asia, such that accusative alignment is mainly found in the east, regardless of language family. Such an 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.

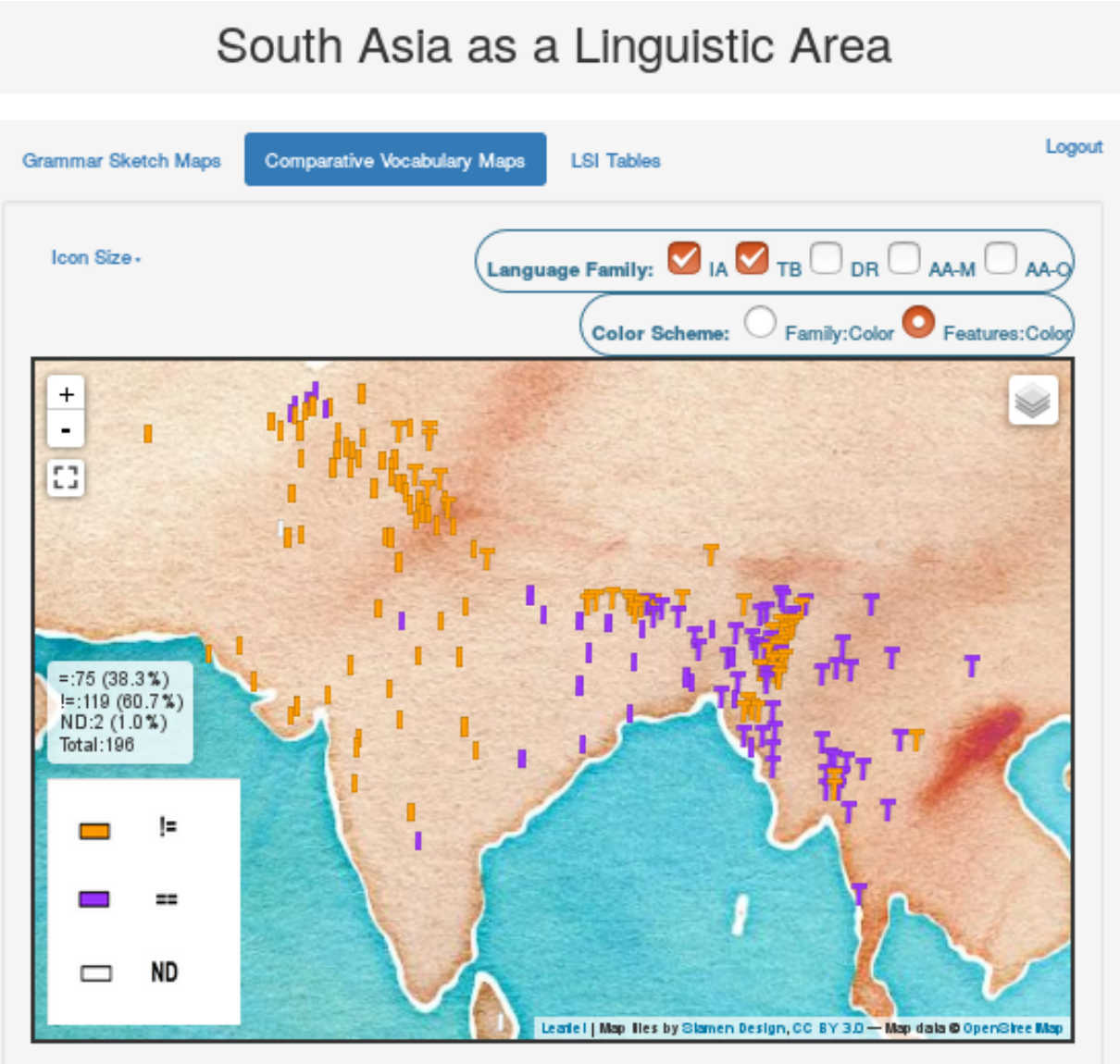

Figure 7: Map showing form of subject pronoun in relation to transitivity

This type of visualization is very helpful for comparison purposes, but not equally useful if we are interested to only explore/visualize feature values of individual languages. For that purpose, we have developed simple feature visualization solutions. Figure 8 shows a screenshot displaying feature values

extracted from the description of Lohorong.[13] This type of expandable/compressible tree styled visualization makes it easy to visualize feature values of a language of interest.

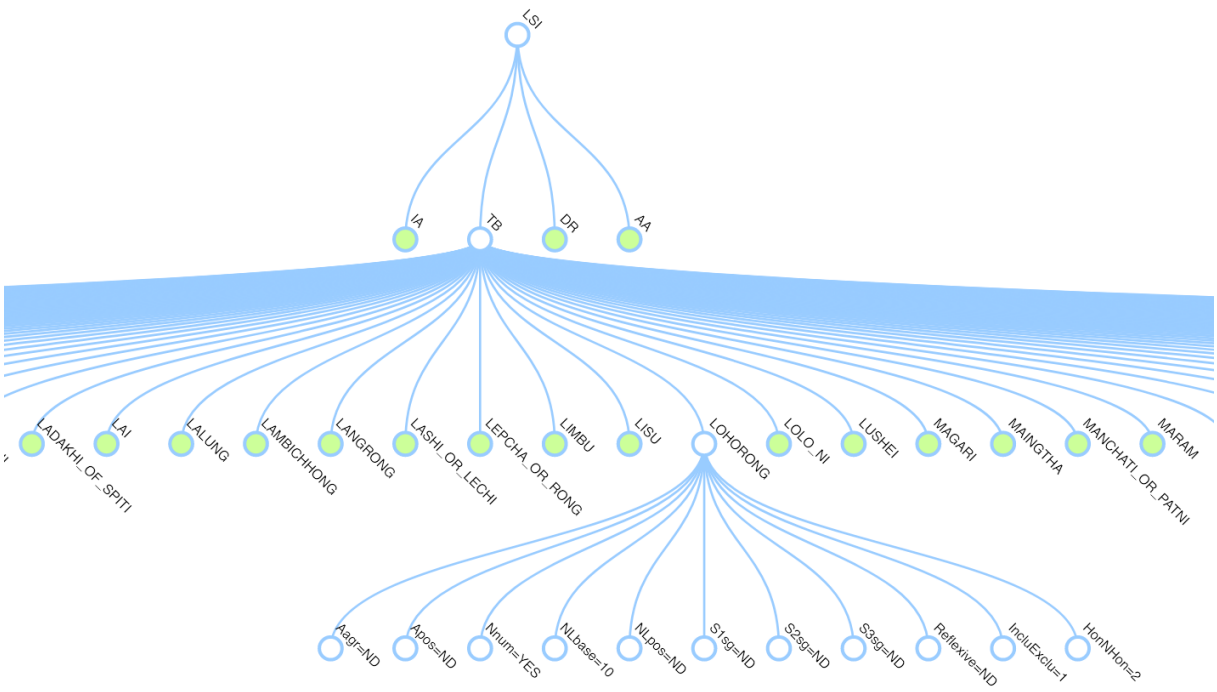

Figure 8: Linguistic features of Lohorong (lrr)

5 Conclusions and future work

Turning the LSI into a structured digital resource will provide a rich empirical foundation for large-scale comparative studies of the linguistic ecology of South Asia. In this day and age, it makes little sense to conduct such studies manually. Instead, they need to draw on extensive digitized language resources and state-of-the-art computational tools. This is the main goal of our ongoing work with the 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. Indeed, the outcome of our experiments on the Grambank data (described in Section 3.3) indicates that there are some grounds for optimism in this regard.

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 7 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

[13] Lohorong (lrr) is a Tibeto-Burman language spoken in Nepal.

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,[14] and the manual questionnaire work and the development of the IE application is underway.

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.

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

[Björkelund et al.2009] Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of CoNLL 2009: Shared Task*, pages 43–48, Boulder, Colorado. ACL.

[Borin and Forsberg2011] Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, pages 41–61. Springer, Berlin.

[Borin et al.2010] Lars Borin, Dana Dannélls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2010. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.

[Borin et al.2012a] 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.

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

[Borin et al.2013a] Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013a. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

[Borin et al.2013b] Lars Borin, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson, and Jonatan Uppström. 2013b. The lexical editing sysem of Karp. In *Proceedings of the eLex 2013 Conference*, pages 503–516, Tallin.

[Borin et al.2014] Lars Borin, Anju Saxena, Taraka Rama, and Bernard Comrie. 2014. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. In *Proceedings of LREC 2014*, pages 3137–3144, Reykjavik. ELRA.

[Chuang et al.2012] 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)*.

[Dryer and Haspelmath2013] Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

[Ebert2006] Karen Ebert. 2006. South Asia as a linguistic area. In Keith Brown, editor, *Encyclopedia of Languages and Linguistics*. Elsevier, Oxford, 2nd edition.

[Evert and Hardie2011] Stefan Evert and Andrew Hardie, 2011. *Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium.* University of Birmingham.

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

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

[Georg2017] Stefan Georg. 2017. Other isolated languages of Asia. In Lyle Campbell, editor, *Language Isolates*, pages 139–161. Routledge, London.

[Grierson1903–1927] George A. Grierson. 1903–1927. *A Linguistic Survey of India*, volume I–XI. Government of India, Central Publication Branch, Calcutta.

[Hammarstedt et al.2017a] Martin Hammarstedt, Lars Borin, Markus Forsberg, Johan Roxendal, Anne Schumacher, and Maria Öhrman. 2017a. Korp 6 – Användarmanual [Korp 6 – User manual]. Research reports from the Department of Swedish GU-ISS 2017-02, University of Gothenburg, Gothenburg. `http://hdl.handle.net/2077/53096`.

[Hammarstedt et al.2017b] Martin Hammarstedt, Johan Roxendal, Maria Öhrman, Lars Borin, Markus Forsberg, and Anne Schumacher. 2017b. Korp 6 – Technical report. Research reports from the Department of Swedish GU-ISS 2017-01, University of Gothenburg, Gothenburg. `http://hdl.handle.net/2077/53095`.

[Havre et al.2000] 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. IEEE.

[Hook1977] Peter E. Hook. 1977. The distribution of the compound verb in the languages of North India and the question of its origin. *International Journal of Dravidian Linguistics*, 6:336–351.

[Janda et al.to appear 2018] Laura A. Janda, Olga Lyashevskaya, Tore Nesset, Ekaterina Rakhilina, and Francis M. Tyers. to appear 2018. A constructicon for Russian: Filling in the gaps. In Benjamin Lyngfelt, Lars Borin, Tiago Timponi Torrent, and Kyoko Hirose Ohara, editors, *Constructicons in Contrast. Constructicography as a Fusion Between Construction Grammar and Lexicography*. John Benjamins, Amsterdam.

[Krstajić et al.2012] 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*, Burlingame, California. SPIE.

[Lyngfelt et al.2012] Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, and Sofia Tingsell. 2012. Adding a constructicon to the Swedish resource network of Språkbanken. In *Proceedings of KONVENS 2012 (LexSem 2012 Workshop)*, pages 452–461, Vienna. ÖGAI.

[Manning et al.2014] 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.

[Masica1976] Colin P. Masica. 1976. *Defining a Linguistic Area: South Asia*. Chicago University Press, Chicago.

[Nichols2003] Johanna Nichols. 2003. Diversity and stability in language. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 283–310. Blackwell, Oxford.

[Saxena2016] Anju Saxena. 2016. Indo-Aryan in typological and areal perspective. Keynote presentation at SALA-32, Lisbon, 27–29 April, 2016.

[Simons and Fennig2018] Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*. SIL International, Dallas, 21st edition. Online version: `http://www.ethnologue.com`.

[Sun et al.2013] 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.

[Surdeanu et al.2003] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, pages 8–15, Sapporo. ACL.

[Swadesh1955] Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

[Ward and Barker2013] Jonathan Stuart Ward and Adam Barker. 2013. Undefined by data: A survey of big data definitions. *CoRR*, abs/1309.5821.

[Xia and Lewis2007] Fei Xia and William Lewis. 2007. Multilingual structural projection across interlinear text. In *Proceedings of HLT 2007*, pages 452–459, Rochester, New York. ACL.