*Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 23–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

# From Linguistic Descriptions to Language Profiles

**Shafqat Mumtaz Virk[1], Harald Hammarström[2], Lars Borin[1], Markus Forsberg[1], Søren Wichmann[3]**
[1]Språkbanken Text, Department of Swedish, University of Gothenburg
[2]Department of Linguistics and Philology, University of Uppsala
[3]Leiden University Centre for Linguistics, Leiden University
[3]Laboratory for Quantitative Linguistics, Kazan Federal University
[3]Beijing Advanced Innovation Center for Language Resources, Beijing Language University
[1]{shafqat.virk, lars.borin, markus.forsberg}@svenska.gu.se
[2]{harald.hammarstrom}@lingfil.uu.se
[3]{wichmannsoeren}@gmail.com

## Abstract

Language catalogues and typological databases are two important types of resources containing different types of knowledge about the world's natural languages. The former provide metadata such as number of speakers, location (in prose descriptions and/or GPS coordinates), language code, literacy, etc., while the latter contain information about a set of structural and functional attributes of languages. Given that both types of resources are developed and later maintained manually, there are practical limits as to the number of languages and the number of features that can be surveyed. We introduce the concept of a *language profile*, which is intended to be a structured representation of various types of knowledge about a natural language extracted semi-automatically from descriptive documents and stored at a central location. It has three major parts: (1) an introductory; (2) an attributive; and (3) a reference part, each containing different types of knowledge about a given natural language. As a case study, we develop and present a language profile of an example language. At this stage, a language profile is an independent entity, but in the future it is envisioned to become part of a network of language profiles connected to each other via various types of relations. Such a representation is expected to be suitable both for humans and machines to read and process for further deeper linguistic analyses and/or comparisons.

**Keywords:** Typological information, linguistic descriptions, language networks

## 1. Introduction

Approximately 7,000 distinct languages constitute our record of linguistic diversity (Hammarström, 2015). Languages are equal witnesses – where e.g., English is but one – to the variation and constraints of the unique communication system of our species (Evans and Levinson, 2009). They harbour information on what happens to language given tens of thousands of millennia of diversification, under all imaginable circumstances of human interaction. As such they may be used to investigate theories that may otherwise not be testable with anything less than a laboratory the size of human history.

Two web-publications maintain catalogues of the languages of the world: Ethnologue (Eberhard et al., 2019) and Glottolog (Hammarström et al., 2019). Ethnologue provides metadata such as number of speakers, location (in prose words), literacy etc. Glottolog provides classification, location (in GPS coordinates), and bibliographical references. For in-depth information about a lesser-known language, specialists typically consult any available descriptive grammar. For example, for the language Ulwa (ISO 639-3 language code: yla) of Papua New Guinea, there exists

> Barlow, Russell. (2018) A grammar of Ulwa. University of Hawai'i at Mānoa doctoral dissertation. xiv+546 pp.

Around 4,000 languages have at least one published grammatical description but the breadth, depth, and quality of these vary (Hammarström et al., 2018).

For analysis of the languages themselves, there are a number of databases which keep a record of various characteristics (also known as linguistic features) of individual languages. For example, the World Atlas of Language Structures (WALS; Haspelmath et al. 2005), contains information on some 200 features spanning 2500 languages (but is sparsely filled in). A very extensive list of linguistic databases can be found at `http://languagegoldmine.com/` (accessed 2020-04-05).

These inventories and databases are highly useful resources but have clear limits on the number of features and/or languages they contain. As such they do not represent all the information available about the same language in descriptive publications. This situation is inevitable as (1) a fixed list of linguistic features is designed for a database, but languages differ from each other in a myriad of ways which cannot be known a priori; and (2) databases are curated manually by reading the descriptive documents, which is a time-consuming activity.

For these reasons we aim to go beyond the manual curation of linguistic databases in order to capture the valuable knowledge about many other languages and features remaining within descriptive publications. Thus, our aim is to extract all the information about a language described in a publication, and represent it in a structured manner. These structured representations can be successively normalized and thus form the basis for large-scale comparison of languages. If successful, it will widen the scope of investigations and comparisons across languages considerably. Toward this end, advancements in natural language processing and information extraction may be exploited.

A related concern is that various types of knowledge about languages are maintained separately. Consequently, one has to explore different resources to access knowledge about the same language. For example, some general and referential type of data (i.e. about language names, the number and names of dialects, the areas where they are spoken, the number of speakers, etc.) are often maintained in the form of digital inventories, the attributive type of data (i.e. various phonological, morphological, and grammatical features) are maintained as typological databases, and many other details are found in descriptive documents (grammars, dictionaries, etc.) and, since recently, increasingly in web-pages, blogs etc.

Further, several of the important resources on natural languages are not open-access. For example, Ethnologue has most of its information behind a paywall.[1] Since only a particular creative arrangement of words – but not facts in general – can be copyrighted, the prospects for free and open structured representations are much better, even when extracted from copyrighted source materials.

In this paper, we present the concept of a language profile in order to address the above-mentioned limitations and concerns. A language profile can be envisaged as a digital representation of a natural language containing various types of information about the language stored at a central location in a structured format and publicly available for further use. It aims to be a dynamic representation in the sense that it is not tied to a predefined set of features (like typological databases), but targets any traceable features. Included are also introductory and referential information about a target language extracted from the descriptions and other available resources. Various types of information about a language are grouped into various sections, and the resulting structure is called *a language profile*. In the present paper, we describe the concept of a language profile only. In future work, we plan to describe how language profiles can be linked in a full network (a LangNet) using different kinds of comparisons/relations (e.g., genetic, geographical, typological similarity). Conceptually, such a network of languages is similar to other networks in the area of NLP such as WordNet, VerbNet, FrameNet, etc., except that it is at the level of languages. We believe that such a rich representation model, and the network of languages will be a useful resource for linguistic studies.

The remainder of the paper is organized as follows: Section 2 describes in detail the structure and components of a language profile, while details on semi-automatic development of a language profile from linguistic descriptions are given in Section 3.

## 2. Language Profiles

As mentioned in the introduction section, a language profile is necessarily a structured digital representation of a natural language. In this section, we will present the structure and various proposed components of a language profile. In doing so, we will use a natural language called ‘Ulwa’, and build a minimal part of its profile. At this stage, this language profile will be built semi-automatically, but a long term objective is to automatize the process as much as possible. We will indicate which parts are built automatically and which manually, and will provide suggestions, wherever possible, for automatizing the corresponding parts.

1. **Metadata Part:** The metadata part contains basic metadata such as official language name, number of speakers, areas where spoken, etc., and referential (e.g., ISO code and/or glottocode, language family, etc.) information. Table 1 shows this part of the ‘Ulwa’ language profile.

   In this case, most of the fields and their values in this part of the profile are available in the language catalogue Ethnologue (Eberhard et al., 2019) in the Yaul entry (`https://www.ethnologue.com/language/yla`). As such it resembles information already available in language inventory databases, but improves on these by being more dynamic, linkable and aggregateable. The list of possible fields of metadata is not bounded, and can be extended indefinitely. Each field in the profile and information within it will have a structured representation. For example, the location in the above given profile is not a simple string, but rather a geographical location with a name and coordinates. This can be linked to existing inventories of geographical locations such as GeoNames (`http://www.geonames.org`). The same applies to the dialect names, families and branches in the classification field, official and alternative language names, etc. Appropriate data structures will be proposed for various fields, with proper IDs to be used for various types of inter- and intra-profile connections. Further, each piece of information will have a recorded source which may be weighted according to usage needs whenever there are many different sources for the same field.

2. **Attributive Part:** This is the major part of a language profile and is intended to contain the typological and other structural information of a target language. Again, other databases exist with a similar type of information (e.g., WALS – see above). The key difference is as follows. The attributive part of a language profile does not contain answers to a predefined set of typological and other questions. Rather, it contains all attributive (i.e. phonological, morphological, and grammatical) information which can be extracted (semi)automatically from the available descriptive data about a given language. As an example, consider the attributive part of the ‘Ulwa’ language profile given in Table 2. The information in this part was automatically extracted from a language description (Barlow, 2018). (A description of the automatic extraction of the typological information is given in Section 3.)

   In this example, there is no categorization of the features. In the future, we intend to divide the attributive part into various subparts e.g. phonological, morphological, grammatical attributive information, and so on. The feature ID field is left blank intentionally at this stage, and a detailed set of feature IDs is to be worked out at a later stage.

[1] Except in third-world countries, where it continues to be freely available.

| Field-ID | Field-Name | Number::Name::Value | Source |
|---|---|---|---|
| fet:p1:meta-name | Official Name(s) | 1::-::Ulwa<br>2:: | (Barlow, 2018) |
| fet:p1:classification | Classification | 1::-::Ulmapo | (Barlow, 2018) |
| fet:p1:speakers | Speakers | 1::Native::700<br>2::Second Language:: | (Barlow, 2018) |
| fet:p1:dialects | Dialect(s) | 1::-::Manu dialect<br>2::-::Maruat-Dimiri-Yaul | (Barlow, 2018)<br>(Barlow, 2018) |
| fet:p1:location | Location(s) | 1::-::Manu<br>2::-::Maruat<br>3::-::Dimiri<br>4::-::Yaul | (Barlow, 2018)<br>(Barlow, 2018)<br>(Barlow, 2018)<br>(Barlow, 2018) |

Table 1: The metadata part of the Ulwa language profile

| **FeatureID** | **Feature** | **Value** | Source |
|---|---|---|---|
| — | Subject and NP order | NP–SubjectMarker | (Barlow, 2018) |
| — | Object and NP order | NP–ObjectMarker | (Barlow, 2018) |
| — | Constituent Order | SOV | (Barlow, 2018) |
| — | PostpositionalPhrase–Oblique-markedNP Order | Both | (Barlow, 2018) |
| — | ObliguePhrase–SubjectOFClause Order | SubjectOFClause-ObliguePhrase | (Barlow, 2018) |
| — | ObliguePhrase–Verb Order | ObliguePhrase–Verb | (Barlow, 2018) |
| — | Negator–Verb Order | Negator–Verb | (Barlow, 2018) |
| — | AdPosition–NP Order | NP–AdPosition | (Barlow, 2018) |
| — | Possessor–Possessum Order | Possessor–Possessum | (Barlow, 2018) |
| — | Adjective–Noun Order | Noun–Adjective | (Barlow, 2018) |
| — | Demonstrative–Noun Order | Noun–Demonstrative | (Barlow, 2018) |
| — | Numeral–Noun Order | Noun–Numeral | (Barlow, 2018) |
| — | RelativeClause–HeadNoun Order | RelativeClause–HeadNoun | (Barlow, 2018) |
| — | PossessivePronoun–Noun Order | PossessivePronoun–Noun | (Barlow, 2018) |
| — | ObliqueMarker–Noun Order | Noun–ObliqueMarker | (Barlow, 2018) |
| — | TraniativeVerb–ObjectMarker Order | TransativeVerb–ObjectMarker | (Barlow, 2018) |
| — | NominalizedVerb–SubjectMarker Order | SubjectMarker–NominalizedVerb | (Barlow, 2018) |
| — | Verb–DirectObject Order | DirectObject–Verb | (Barlow, 2018) |
| — | TransitiveVerb–ObjectMarker Order | ObjectMarker–TransitiveVerb | (Barlow, 2018) |
| — | Oblique–Verb Order | Oblique–Verb | (Barlow, 2018) |
| — | Oblique- Subject Order | Subject–Oblique | (Barlow, 2018) |
| — | Adverb–Subject Order | Subject–Adverb | (Barlow, 2018) |
| — | Adverb–Object Order | Adverb–Object | (Barlow, 2018) |
| — | Adverb–Oblique-markedNP Order | Adverb–Oblique-markedNPs | (Barlow, 2018) |
| — | NasalSegments–VoicelessStops Order | NasalSegments–VoicelessStops | (Barlow, 2018) |
| — | LabialNasal–PalatoAlveolar Order | LabialNasal–PalatoAlveolar | (Barlow, 2018) |
| — | HomorganicNasals–VoicelessStops Order | HomorganicNasals–VoicelessStops | (Barlow, 2018) |
| — | Liquids–LabialStops Order | LabialStops–Liquids | (Barlow, 2018) |
| — | Liquids–VelarStops Order | VelarStops–Liquids | (Barlow, 2018) |

Table 2: The attributive part of the Ulwa language profile

3. **References Part:** The reference part contains a list of available resources about the language at hand. A BibTeX type of entry will be maintained for each descriptive document and other type of resource (e.g. word list, dictionary, etc.). One such entry for the ‘Ulwa’ language is as follows:

```
        @phdthesis{g:Barlow:Ulwa,
        author = {Barlow, Russell},
        title = {A grammar of Ulwa},
        school = {University of Hawai’i
                  at Mānoa},
        pages = {xiv+546},
        year = {2018},
        glottolog_ref_id = {554079},
        hhtype = {grammar},
        inlg = {English [eng]},
        lgcode = {Manu Ulwa = Yaul [yla]},
        macro_area = {Papua}
    }
```

Every item of information in each section of the language profile has a source linked to an entry from the reference section. The maintenance of references within the profile ensures that the crucial source links can be kept in sync.

## 3. Building a Language Profile

Building a language profile is a complex process. It requires gathering information about a language from all available sources, i.e., manuals, digital inventories, linguistic descriptions, etc. This is a long-term process, and will require gradual efforts to incrementally develop a large set of rich language profiles.

At this stage, we have relied on manual collection of information for the introductory as well as the reference part, although parts of it can be automatized (information about language name and number of speakers can be extracted automatically using the frame based methodology explained below, which was used to build the attributive part automatically).

The automatic extraction of typological information from descriptive grammars is a novel task, and there exists only a few studies and systems reported previously (Virk et al., 2017; Virk et al., 2019). In Virk et al. (2019), a frame-semantic based approach is proposed for developing a parser to automatically extract typological linguistic information from plain-text grammatical descriptions of natural languages. As a case study, the authors have shown how the parser can be used to extract value of an example typological feature. However, the system has not been used for any actual typological work. We continue that work and use the parser to extract typological feature values (shown in Table 2) of a language profile. A brief description of the parser and how it has been used for our purposes follows.

The parser relies on a lexico-semantic resource, LingFN (Malm et al., 2018), and its frame-labeled data for training machine learning models to build a parser. The development of LingFN itself is based on the theory of frame-semantics (Fillmore, 1976; Fillmore, 1977; Fillmore, 1982), and is motivated by the development of Berkeley FrameNet (Baker et al., 1998) and other, domain-specific framenets (e.g. a framenet to cover medical terminology (Borin et al., 2007), *Kicktionary*,[2] a soccer language framenet). Let us take an example to better understand what LingFN is, and how its frame-labeled data is used to build the frame-semantic parser which in turn is used for automatic extraction of typological features. Consider the following sentence which is taken from a descriptive grammar of the Ulwa language.

> In Ulwa, adjectives in NPs sometimes precede their head nouns.

The sentence contains information about the relative position (sequencing) of two syntactic categories i.e. ‘adjectives’ and ‘head nouns’. Their position wrt one another is not always the mentioned one but could be the other way around, as conveyed by the adverb ’sometimes’. This is useful information that we are interested in extracting automatically. One of the possible approaches is to develop a frame-semantic based information extraction system. For that purpose, the first step is to design (or use from the Berkeley FrameNet) special structures to represent this type of phenomenon (i.e. sequencing). In frame-semantics such structures are called semantic frames, and in general, a semantic frame is a structured representation of a an entity, an object, and a scenario. In our case, a semantic frame represents a linguistic entity (e.g. nouns, verbs, etc.) or phenomenon (e.g. affixation, agreement, sequence, etc). Let us say for the sequencing phenomena, we have designed a semantic frame with the structure shown in Table 3. (For more details on development of the SEQUENCE and other linguistic domain semantic frames with annotated example sentences, we refer the reader to Malm et al. (2018)).

| SEQUENCE |
|---|
| Entity_1 |
| Entity_2 |
| Entities_3 |
| Order |
| Frequency |
| Language_Variety |

Table 3: Sequence Semantic Frame

Entity_1, Entity_2, Entities, Order, Frequency, and Language_Variety are referred to as frame-elements, which constitute various semantic parts of the sequencing phenomena. With such structures (i.e. semantic frames) at hand, the next step is to annotate linguistic descriptions with developed semantic frames. The annotation of the above given sentence is as follows:

```
In [Ulwa]_Language Variety,
   [adjectives]_Entity_1 in NPs
   [sometimes]_Frequency
   [[precede]_LU]_Order
   their [head nouns]_Entity_2.
```

[2] http://www.kicktionary.de/

String segments labeled as one of the frame-elements are enclosed within a pair of brackets while the frame-element label (bold) follows an underscore sign. Note that in case of above given sentence, the word 'precede' is both a lexical unit (a word triggering a particular frame) and also a frame-element.

Now imagine, if we have enough sentences annotated with the SEQUENCE (and other frames from LingFN), one could train machine learning models for automatic labeling of these frames on un-annotated data. This is exactly what is proposed by the authors in (Virk et al., 2019), and they have a developed a parser for this purpose. What the parser does is to take un-annotated sentences containing typological linguistic information and annotate them with linguistic domain frames and their associated frame-elements. As suggested by the authors in the same paper, the annotations can be converted to typological information in any required format using a rule based module. This is exactly what we have done to extract feature values shown in Table 2 for the Ulwa language. Note, only the SEQUENCE frame was used to extract the whole information present in Table 2. In the future we plan to extend this work to other typological features and hence enhance the attributive part of a language profile.

## 4. Conclusions and Future Work

We have presented the idea of a language profile, which is envisaged as a digital structured representation of a natural language. It has two major objectives. The first objective is to overcome a major limitation of existing typological databases which contain information about a pre-defined set of linguistic features. We propose work towards automatically extracting information about all the features described in a descriptive document. The second objective is to collect various types of information available about a language stored in a structured way and at a common place together with information about the sources. The idea is at an embryonic stage and is to be further matured and extended in the future.

## 5. Acknowledgements

The work presented here was funded by (1) the *Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage* (DReaM) Project awarded 2018–2020 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden; (2) 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); (3) *From Dust to Dawn: Multilingual Grammar Extraction from Grammars* project funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond 2007.0105, Uppsala University; and (4) the University of Gothenburg, its Faculty of Humanities and its Department of Swedish, through their truly long-term support of the Språkbanken Text research infrastructure.

## 6. References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of ACL/COLING 1998*, pages 86–90, Montreal. ACL.

Barlow, R. (2018). *A grammar of Ulwa*. Ph.D. thesis, University of Hawai'i at Mānoa.

Borin, L., Toporowska Gronostaj, M., and Kokkinakis, D. (2007). Medical frames as target and tool. In *FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages. (Nodalida 2007 Workshop Proceedings)*, pages 11–18, Tartu. NEALT.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). *Ethnologue: Languages of the World*. Dallas: SIL International, 22 edition.

Evans, N. and Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, pages 55–81. North Holland, Amsterdam.

Fillmore, C. J. (1982). Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul.

Hammarström, H., Castermans, T., Forkel, R., Verbeek, K., Westenberg, M. A., and Speckmann, B. (2018). Simultaneous visualization of language endangerment and language description. *Language Documentation & Conservation*, 12:359–392.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). Glottolog 4.0. Jena: Max Planck Institute for the Science of Human History. Available at `http://glottolog.org`. Accessed on 2019-09-12.

Hammarström, H. (2015). Ethnologue 16/17/18th editions: A comprehensive review. *Language*, 91(3):723–737. Plus 188pp online appendix.

Martin Haspelmath, et al., editors. (2005). *World Atlas of Language Structures*. Oxford: Oxford University Press.

Malm, P., Virk, S. M., Borin, L., and Saxena, A. (2018). Lingfn: Towards a domain-specific linguistic framenet. In Tiago Timponi Torrent, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Virk, S., Borin, L., Saxena, A., and Hammarström, H. (2017). Automatic extraction of typological linguistic features from descriptive grammars. In *Proceedings of TSD 2017*, pages 111–119, Cham. Springer.

Virk, S. M., Muhammad, A. S., Borin, L., Aslam, M. I., Iqbal, S., and Khurram, N. (2019). Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. In *RANLP-Proceedings*, sep.