The Alaskan Athabascan Grammar Database

Sebastian Nordhoff, Siri Tuttle, Olga Lovick
Glottotopia, Alaska Native Language Center, First Nations University of Canada
Berlin, Fairbanks, Regina
sebastian.nordhoff@glottotopia.de, sgtuttle@alaska.edu, olga@lithophile.com

Abstract

This paper describes a repository of example sentences in three endangered Athabascan languages: Koyukon, Upper Tanana, Lower Tanana. The repository allows researchers or language teachers to browse the example sentence corpus to either investigate the languages or to prepare teaching materials.

The originally heterogeneous text collection was imported into a SOLR store via the POIO bridge. This paper describes the requirements, implementation, advantages and drawbacks of this approach and discusses the potential to apply it for other languages of the Athabascan family or beyond.

Keywords: Athabascan languages, linguistic examples, least-resourced languages

1. Introduction

This paper presents the Alaskan Athabascan Grammar Database (AAGD).

The goal of this database is to make available comparable annotated grammar examples from all eleven Athabascan languages, drawn from texts, lexicons, grey literature and new fieldwork, accessible to researchers and community members. This poses a number of technical and conceptual challenges, which will be explored in this paper.

2. Athabascan languages

The Athabascan language family numbers comprises over 40 languages (Mithun, 1999). Of those 11 are spoken in Alaska, with a total number of speakers of about 5300. These are: Ahtna, Deg Xinag, Denaina/Tanaina, Gwichin, Hän, Holikachuk, **Koyukon**, **Lower Tanana**, Middle Tanana, Tanacross, **Upper Tanana**, Upper Kuskokwim. Languages in bold are included in the first phase of the project.

As far as the orthography is concerned, the languages are written in the standard Latin alphabet with the following special characters: accents for tone, ogoneks for nasalization, <ł> for the lateral fricative, and diacritics and digraphs marking language-specific contrasts.

Standardization of orthographic conventions is not complete for two of the languages, and both phonemic inventories and orthographies vary considerably, so that automated generalizations across languages are difficult to achieve.

Athabascan languages are known for their complex morphology. An example is given in (1).

(1) *Neeghonoheedekkaanh.*

Nee-	*gho-*	*no-*	*h-*	*∅-*	*ee-*
TERM-	PPOST-	REV-	3PL.SBJ-	∅.CNJ-	PFV-
de-	*kkaanh*				
D.CLF-	paddle.PFV				

'They paddle back to shore.' (Koyukon)

(1) illustrates that *a fully inflected Athabascan verb form can convey a complete utterance*. This could be taken to suggest that verbs are isomorphic with sentences, and that there is no syntax outside the verb. While this is not a serious claim, it is still the case that syntactic research in Athabascan languages considerably lags behind that on morphology cf. (Rice, 2000, 1) and phonology.

3. Use case

The long term goal of this project is to make all existing textual data from the Athabascan languages in Alaska digitally available for (syntactic) research, complemented by new field data for the languages where data is lacking and where collection is still possible. There is a vast amount of grey literature for those languages (unpublished theses, manuscripts, legacy computer files on researchers' computers). Making this accessible to researchers and language learners is also a goal.

For the initial phase, three languages were chosen: Upper Tanana, Lower Tanana, and Koyukon. These languages were chosen because of current research activity by Tuttle and Lovick as well as because of the quality of the available material and the possibility to conduct further fieldwork.

As for the linguistic scope, a focus is on syntax, as this has been a neglected area of research for these languages in the past. This decision has no direct consequences for data collection, but it does influence data annotation and supplementary pedagogical material to be provided. Furthermore, it means that morphological annotation need not be as detailed. This is crucial, as deep morphological annotation of the kind exemplified in (1) is extremely time-consuming.

The project is special in that it has two audiences: on the one hand academic researchers in linguistics, who can be assumed to have the required background in terminology and linguistic theory to make sense of the structure of the data provided. On the other hand there are language teachers, who might have very good or reasonable knowledge of the language, but lack the training in language description to appreciate the difference between clitics and affixes for instance. Catering to these two audiences at once remains a challenge.

The main use case for the academic audience is a repository of sentences with good search functionalities. Next to string search, search by tags/categories (e.g. "contains past tense" or "contains negation") and search by similarity should be provided.

A further use case is the creation of a model for grammatical comparison transferable to other, unrelated, language families.

For the language teacher audience, the main use case is the preparation of lessons. Here, access to examples particularly suited to illustrate a certain point ("Exemplars" (Good, 2004)) is crucial . Teachers preparing a lesson on negation, for instance, should have access to relevant sentences illustrating the phenomenon under discussion. Ideally, the sentences should be sorted according to their accessibility: straightforward sentences should be separated from contrived ones presenting difficulties unrelated to the phenomenon to be addressed.

4. Data collection

Data in our project is drawn from archival sources, published sources and new fieldwork. The three types of data have to be dealt with differently to "collect" them and prepare them for use in the database. The eleven Athabascan languages of Alaska vary in the amount and type of data that has been archived. The Alaska Native Language Archive at `http://www.uaf.edu/anla` lists the following number of records:

- Ahtna 187
- Deg Xinag 205
- Denaina 400
- Gwichin 689
- Hän 120
- Holikachuk 72
- **Koyukon** 532
- **Lower Tanana** and Middle Tanana 225
- Tanacross 170
- **Upper Tanana** 149
- Upper Kuskokwim 155

These numbers are inflated, however, as they also include materials *about* the language, not *in* the language. The actual number of usable texts is much smaller.

Archival sources include .pdfs of original notes, .pdfs of typescripts of notes and texts, audio and video recordings of various activities, and .pdfs of gray material such as lesson books and classroom materials, among others. There is great variety in the applicability of these materials, as well as their completeness in terms of glossing, analysis and metadata. These materials cannot be imported directly into the project without annotation.

Published sources include dictionaries and text collections. In several cases, digital versions exist and are archived (for (Jetté and Jones, 2000), for example). However, because of lacking metadata and shallow analysis, annotation is needed for these materials as well.

New fieldwork also produced data of different types. While new texts have been collected, annotated and uploaded during the project period, other field work has concentrated on preparation of archival texts, completing translation and analysis and adding metadata as well as on the elicitation of additional grammatical data using a variety of stimuli. All new material, recorded with audio or video, requires transcription, translation, and annotation.

5. Requirements

The requirements for the software were identified as the following:

- all three languages should be represented;
- it should be possible to extend the platform to the other (Alaskan) Athabascan languages, and potentially any other language;
- there should be a generic way to import data in various linguistic formats. (ELAN,[1] Toolbox,[2] Typecraft,[3] Brat[4])
- there should be a way to annotate the data after it is imported;
- there should be a way to retrieve the data;
    - full text search;
    - category search;
    - similarity search;
- the platform should be usable for researchers;
- the platform should be usable for language teachers;
- it should be possible to single out certain examples as particularly well suited for a certain didactic point;
- it should be possible to correct minor errors online;
- users should have the possibility to upload additional texts;
- user management and security;
- possibility to add prose texts explaining certain phenomena.

6. Implementation

6.1. Import

The project uses a toolchain with POIO (Bouda et al., 2012)[5] as a hub (Figure 1). POIO takes a variety of input formats, among which the ELAN format, which is now widespread in language documentation projects. POIO transforms all those input formats in to LAF/GrAF (Ide and Romary, 2006; Ide and Suderman, 2007). This allows us to be agnostic of the actual input format and focus on the conversion of LAF/GrAF into an XML format to import into SOLR.[6]

[1] `https://tla.mpi.nl/tools/tla-tools/elan/`

[2] `http://www-01.sil.org/computing/toolbox/`

[3] `http://typecraft.org/tc2wiki/Main_Page`

[4] `http://brat.nlplab.org/`

[5] `https://github.com/cidles/poio-api`

[6] `http://lucene.apache.org/solr/`

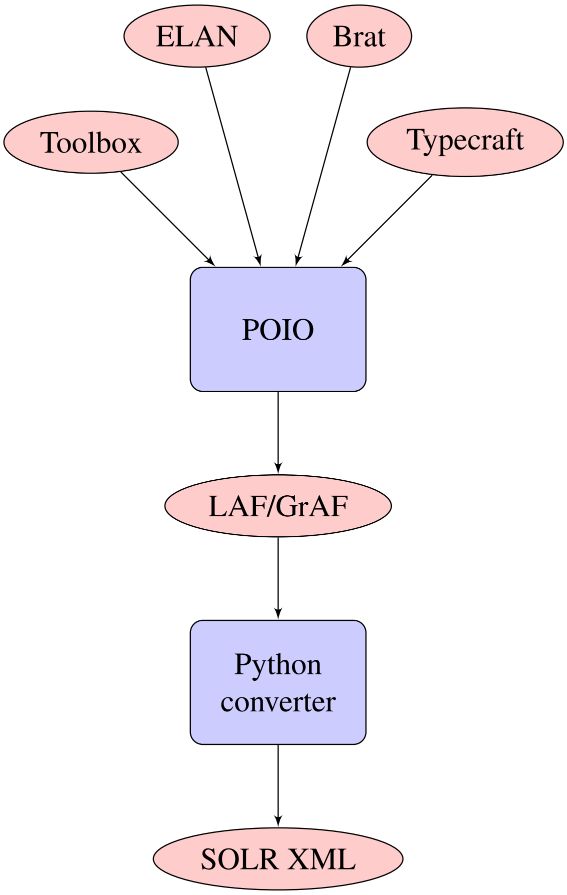
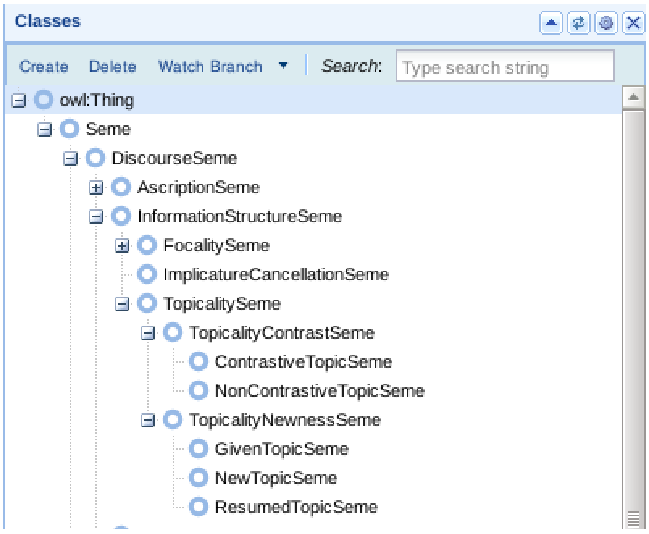


Figure 1: The conversion tool chain. Files are give in red, programs are given in blue. Import of SOLR XML into SOLR store not shown.

We use a SOLR setup with only minimal changes to the schema and configuration provided in the example installation. In particular, we add 20 fields for the domains we use for annotation.

An important design choice was the representation of interlinearized glossed text. ELAN comes with a configurable hierarchy of named tiers. There is a fixed level of named tiers of linguistic organization (morphemes, words, clauses, utterances). The larger units are subdivided into units of the next lower type (e.g. clauses are subdivided into words). At each level of organization, the items can be annotated (typical: translation, parts-of-speech). Tiers can have different properties and different relations to each other. Rather than following the elaborate ELAN model, however, we opted for a much simpler meronymic approach of unnamed and untyped tiers. We use a recursive nesting structure to represent the part-whole relationships. There is a generic XML element `item` which can have the child items `label`, `translation`, `pos` (part-of-speech) and `children`. `children` is a list of further `items`.

Figure 3 shows how the relation between vernacular text and translation/gloss is rendered uniformly across the sentence level, the word level and the morpheme level (white for vernacular, grey for translation/gloss). The relation between the larger blocks and the smaller blocks they contain is also rendered uniformly. If the number of levels changed, the representation would expand/reduce accordingly.



owl:Thing
- Seme
  - DiscourseSeme
    - AscriptionSeme
    - InformationStructureSeme
      - FocalitySeme
      - ImplicatureCancellationSeme
      - TopicalitySeme
        - TopicalityContrastSeme
          - ContrastiveTopicSeme
          - NonContrastiveTopicSeme
        - TopicalityNewnessSeme
          - GivenTopicSeme
          - NewTopicSeme
          - ResumedTopicSeme

Figure 2: Information structure in OCCULT.

6.2. Annotation

Annotation is done by the linguists in the web frontend. After experimenting with the extended OCCULT ontology (990 concepts),[7] the linguists found it easier to settle for a smaller shallow domain specific ontology consisting of 180 concepts. The following list shows the shallow ontology for information structure while Figure 2 shows the relevant excerpt of OCCULT. Among the differences we find: 1) the higher nodes of the ontology (Seme>Discourse Seme) are not displayed and 2) the two separate concepts of contrastiveness and newness are merged and the cross-product is given as a flat list. This allows for faster annotation compared to selecting annotations in a more articulate tree.

- information_structure
  - non-contrastive_given
  - non-contrastive_resumed
  - non-contrastive_new
  - contrastive_given
  - contrastive_resumed
  - contrastive_new

There is a clear separation between formal concepts like `sentence_type:monoclausal` and meaning-based concepts like `participant_role:experiencer`, following (Nordhoff, 2012).

Annotation is done via AJAX and a small python script on the server, which validates the input, updates the SOLR store and returns the callback message.

6.3. View

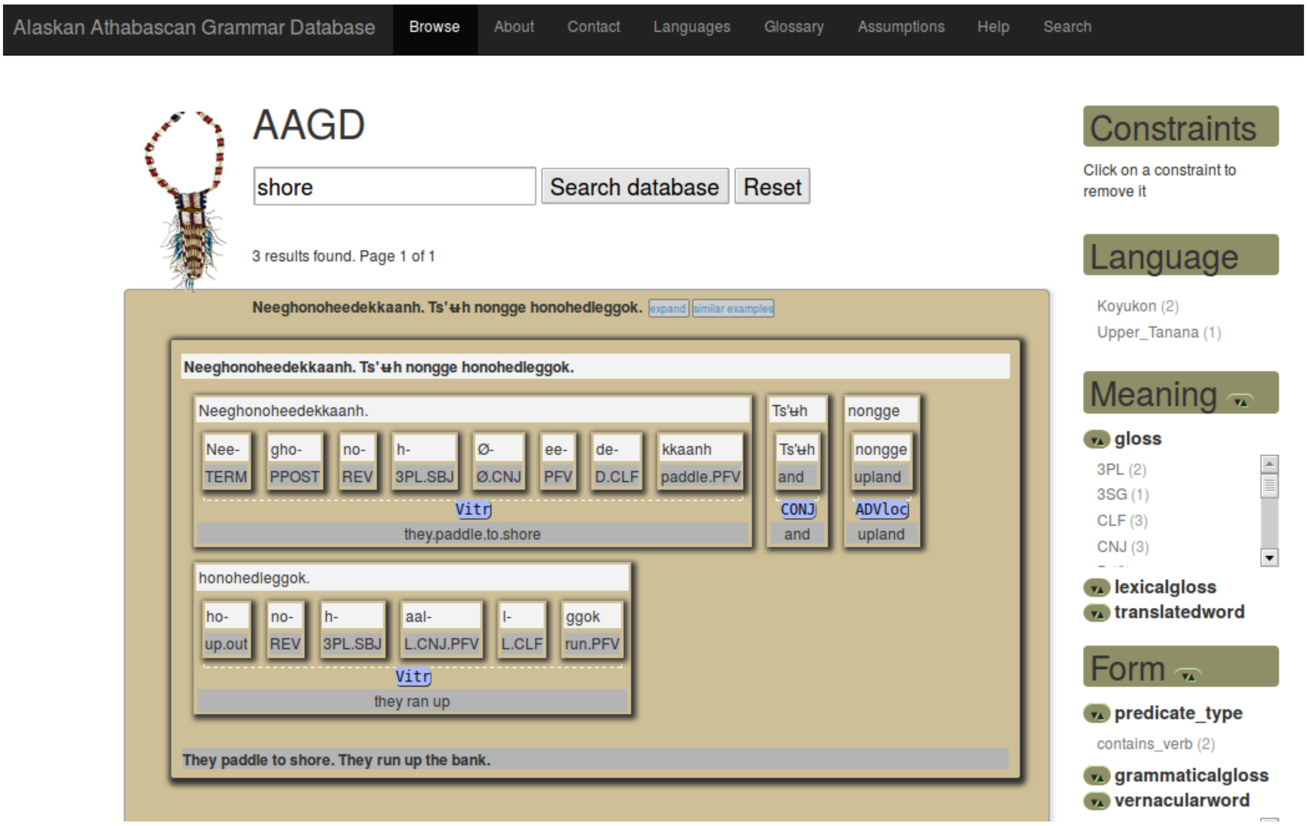
The velocity[8] templates shipped with SOLR were completely rewritten, using JQuery[9] and Bootstrap.[10] A Moin-

[7] https://github.com/Glottotopia/ontologies/tree/master/occult

[8] http://velocity.apache.org/

[9] http://jquery.com/

[10] http://getbootstrap.com/



shore

3 results found. Page 1 of 1

Neeghonoheedekkaanh. Ts'ʉh nongge honohedleggok.

Neeghonoheedekkaanh.

Nee-	gho-	no-	h-	Ø-	ee-	de-	kkaanh
TERM	PPOST	REV	3PL.SBJ	Ø.CNJ	PFV	D.CLF	paddle.PFV

Vitr
they.paddle.to.shore

Ts'ʉh

Ts'ʉh
and

CONJ
and

nongge

nongge
upland

ADVloc
upland

honohedleggok.

ho-	no-	h-	aal-	l-	ggok
up.out	REV	3PL.SBJ	L.CNJ.PFV	L.CLF	run.PFV

Vitr
they ran up

They paddle to shore. They run up the bank.

Koyukon (2)
Upper_Tanana (1)

3PL (2)
3SG (1)
CLF (3)
CNJ (3)

contains_verb (2)

Figure 3: Screenshot of the web frontend with a free text search, one fully visible example in the middle and faceted search on the right

moin[11] wiki is used for documentation. Moinmoin also serves as the didactic frontend for the SOLR store. Here, explanatory chapters can be written. Exemplars for the illustration of the content of the prose text can be pulled from the store via an AJAX request. The JSON data returned are processed with javascript and integrated into the HTML of the wikipage. The Moinmoin anchor notation (not used elsewhere on the installation) serves as an easy way for end users to pull examples from the SOLR-store. `||<tableclass="idselector"> ||<<Anchor(UTOLVDN07Aug2205-053)>>||` in the wiki markup of the page will result in the example with the ID UTOLVDN07Aug2205-053 being pulled from the SOLR store and displayed in the HTML.

Didactic usefulness can be divided into three categories. The first one are singular, cherry-picked examples which are discussed in the prose text with their particular properties, referring to particular words or morphemes of the text, their order and other morphosyntactic or semantic properties thereof. These examples have to be identified by ID. The second layer consists of the type "Further reading". These examples are not discussed individually, but are nevertheless useful illustrations of the phenomenon at hand. These can be identified by tag. The third layer are all examples which involve the phenomenon at hand no matter whether it is central in the example or whether the example is involved, overlong, incomplete or presents other hurdles to didactic use. Examples of this third layer would typically not be pulled into the wiki pages but are of course still accessible via the SOLR store.

7. Justification for this implementation

The POIO hub allows for a uniform treatment of linguistic data regardless of actual input format. The SOLR store brings Lucene search capabilities out of the box and relieves us from the need to provide the search facilities ourselves. Given the rather small amount of data, there is no need to use a RDBMS in the backend; the SOLR XML store is sufficient for the amount of data handled here. Another strong argument were the search capabilities SOLR offers out of the box. True, one could use a relational database and build a SOLR index just for querying, but the amount of data we are dealing with in Athabascan linguistics is not large enough for the difference in performance to be noticeable.

Some of the data are stored in a denormalized way. For instance, the examples are stored as one big XML-string and the translation is again stored as a string for search purposes. In a relational world, one would store the components of a linguistic example (source, interlinear gloss, translation) in a granular fashion and then reference the relevant fields to reconstitute the example. This would avoid mismatches between the translation in the big XML-string and the translation in the separate field, which could occur with the current implementation. However, a full modeling of the dependencies of the elements of linguistic examples is not required for our use case, which is first and foremost a syntactic one.

[11] `http://moinmo.in/`

The shallow ontology is a compromise between a more articulate ontology and user experience when annotating. When experimenting with the more articulate ontology, we found that deep-nested categories reduced efficiency in annotation due to the time needed to navigate the tree. We also used free text search fields to quickly find a desired concept by name. This, however, requires that the users really know the ontology well, which cannot be assumed as a given.

8. Availability

The source code is available on github at `https://github.com/Glottotopia/aagd`. The website is available at `http://www.glottotopia.org/aagd`. The language data are still being curated and will be made available in due course.

9. Significance

Athabascan languages present least-resourced languages. The tools typically found in LREC like treebanks, thesauri, wordnets and the like are far away for languages where even a full grammar and a comprehensive dictionary are lacking. Nevertheless, there are resources in these languages, and modern technological tools developed for other languages like LAF/GrAF or SOLR can be fruitfully applied to these languages, even if the use cases are very different. This requires a lot more massaging than with larger languages. Lack of a standardized orthography, for instance, means that during import choices have to be made as to the string representation. Possibly, two competing orthographies in the input documents cover different aspects of the phonology of a language, and it is unclear how one could be translated into the other without access to native speakers. The dearth of data and the need for manual inspection of all utterances entail that the amounts of data available are tiny compared to larger languages. Furthermore, in order to start automated syntactic processing of these languages, one has to get a clearer picture of these languages' syntax in the first place. This project is thus a very small step into the direction of automated processing of Athabascan linguistic data: allowing humans to browse the data.

On the other hand, languages with a very different typology from the languages typically dealt with in LREC (mostly Indo-European languages or larger East Asian languages) are a nice test case for existing technologies. For the time being, we have only shown that the POIO bridge can be applied to these languages, and that the data can usefully be stored in and retrieved from a SOLR store. Neither of these technologies has a very important linguistic apparatus built into its design; Still, it is a first step towards more adept use of existing computational technologies for least-resourced languages. Since many documentation projects of endangered languages use the ELAN file format, the POIO bridge has a good potential of importing these resources into a SOLR store with reasonable adapatation.

We have furthermore started porting the project to a corpus of 10,000 grammatical descriptions, which we have mined for linguistic examples. First results show that the approach and querying facilities scale to a world-wide level with examples from several 1000 languages. This will remain at the proof-of-concept level, though, due to copyright restrictions for the grammatical descriptions used. It is hoped that in the future, more linguistic data will be made available under a license which allows reuse. See (Schenner and Nordhoff, 2016) for an example of mining linguistic examples from Open Access books by Language Science Press.

10. Acknowledgments

This project was supported by the NSF grant NSF/BCS 1160654: Alaskan Athabascan Grammar Database Development, $120,521.00. SG Tuttle, PI, Olga Lovick, Co-PI.

11. Bibliographical References

Bouda, P., Ferreira, V., and Lopes, A. (2012). Poio API - an annotation framework to bridge language documentation and natural language processing. Paper presented at The Second Workshop on Annotation of Corpora for Research in the Humanities.

Good, J. (2004). The descriptive grammar as a (meta)database. Paper presented at the EMELD Language Digitization Project Conference 2004. http://linguistlist.org/emeld/workshop/2004/jcgood-paper.html.

Ide, N. and Romary, L. (2006). Representing linguistic corpora and their annotations. In *Proceedings of the 5th Language Resources and Evaluation Conference*. Genoa/Italy.

Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop, 18, Prague/Czech Republic*.

Jetté, J. and Jones, E. (2000). *Koyukon Athabaskan Dictionary*. Alaska Native Language Center, Fairbanks.

Mithun, M. (1999). *The Languages of Native North America*. Cambridge University Press, Cambridge.

Nordhoff, S. (2012). The grammatical description as a collection of form-meaning pairs. In Sebastian Nordhoff, editor, *Electronic Grammaticography*, pages 33–62. University of Hawai'i Press, Manoa.

Rice, K. (2000). *Morpheme Order and Semantic Scope*. Cambridge University Press, Cambridge.

Schenner, M. and Nordhoff, S. (2016). Extracting interlinear glossed text from LaTeX documents. In *LREC 2016*.