Towards LLOD-based Language Contact Studies: A Case Study in Interoperability

Christian Chiarcos, Kathrin Donandt, Hasmik Sargsian, Jesse Wichers Schreur, Maxim Ionov
Goethe University Frankfurt, Frankfurt am Main, Germany
{chiarcos|donandt|ionov}@cs.uni-frankfurt.de,
{sargsyan|wichersSchreur}@em.uni-frankfurt.de

Abstract

We describe a methodological and technical framework for conducting qualitative and quantitative studies of linguistic research questions over diverse and heterogeneous data sources such as corpora and elicitations.
We demonstrate how LLOD formalisms can be employed to develop extraction pipelines for features and linguistic examples from corpora and collections of interlinear glossed text, and furthermore, how SPARQL UPDATE can be employed
(1) to normalize diverse data against a reference data model (here, POWLA),
(2) to harmonize annotation vocabularies by reference to terminology repositories (here, OLiA),
(3) to extract examples from these normalized data structures regardless of their origin, and
(4) to implement this extraction routine in a tool-independent manner for different languages with different annotation schemes.
We demonstrate our approach for language contact studies for genetically unrelated, but neighboring languages from the Caucasus area, Eastern Armenian and Georgian.

Keywords: Linguistic Linked Open Data, language contact, Georgian, Armenian, syntax, corpus interoperability

1. Motivation

We describe a methodological and technical framework for qualitative and quantitative investigations of linguistic research questions which heavily depend on data such as corpora, elicitations, etc. It can be used for all research areas, but is primarily suitable for typological, historical and comparative studies. We demonstrate our approach using a specific research question in language contact studies as a case study.

For such research, there are usually several data sources, e.g. a dictionary, a number of elicitations, or even a corpus. All of these may be in different formats without an interface to query over them simultaneously. Furthermore, these linguistic resources may not even share a tagset, and may have different annotations for the same grammatical categories. We show that by applying (Linguistic) Linked Open Data (LLOD) principles, we are able to unify different types of resources, and query these heterogeneous sources as a single united resource.

Linguistic Linked Open Data (LLOD)[1] describes the application of Linked Open Data principles and methodologies for modeling, sharing and linking language resources in various text- and knowledge-processing disciplines. These disciplines range from artificial intelligence and computational linguistics via lexicography and the localization industry to linguistics and philology. For these areas, a number of benefits of LLOD and the underlying RDF technology over traditional representation formalisms have been identified (Chiarcos et al., 2013). Most notable for the work described here, this includes *representation* (linked graphs can represent any kind of linguistic annotation), *interoperability* (RDF graphs can easily be integrated), *ecosystem* (broad support by off-the-shelf database technology), and explicit *semantics* (links to commonly used vocabularies provide community-approved meanings for concepts and data structures).

LOD interoperability and the ability to use its shared vocabularies provides the possibility to integrate and enrich different and heterogeneous language resources. In our project, we focus on applying this methodology to studies in various areas of linguistics: Armenian and Kartvelian studies, language contact studies, syntax, and typology.

In this paper, we show the application of this approach on the study of similar syntactic constructions in Standard Eastern Armenian and Modern Georgian using heterogeneous resources. In order to use those resources we convert them to a unified representation. Using RDF conversion and further SPARQL UPDATE queries, we create a pipeline that dynamically annotates a data stream (with a help of `CoNLLStreamExtractor`, a part of the CoNLL-RDF library (Chiarcos and Fäth, 2017)[2]). The enriched annotation can then be used to conduct the research at hand.

The remainder of the paper is structured as follows: Section 2 introduces the linguistic problem under consideration, Section 3 presents the corpus data and explains its conversion to a unified format which is a necessary preparation step for the experiment described in Section 4. Section 5 reflects on the results of the experiment and the insights gained, and discusses its relevance for approaching the linguistic problem at hand.

2. Linguistic Background

2.1. Introduction

Georgian and Armenian are genealogically unrelated languages that have been spoken in neighboring areas for centuries. Hence, they are expected to share a number of features on different levels of linguistic analysis, among which

[1] http://linguistic-lod.org/

[2] https://github.com/acoli-repo/conll-rdf

syntax. One of the common syntactic-pragmatic features of Georgian and Armenian is pre-verbal focus (Comrie (1984, pp.1-2); Harris (1981, pp.14-18)). With pre-verbal, we mean the position directly before the finite verb, which can be either a main verb or an auxiliary. See Section 2.2 for a short discussion of focus.

As a case study, we look into common analytic predicative constructions in these languages, namely those that consist of an auxiliary verb and a main verb. More specifically, we consider the position of the auxiliary with respect to the main verb. This will serve as a basis for a further research on the effects of word order on the focus of the clause. If the results are similar in both languages, this would be a possible testament to syntactic convergence in the history of these neighboring languages.

We restrict our preliminary research on word order samples to the *to-be*-auxiliary and a modal auxiliary in Armenian, and three modal auxiliaries in Georgian.

2.2. Terminology

There is hardly a completely unambiguous and cross-linguistically valid definition for the term 'auxiliary (verb)' (Ramat, 1987, pp. 3-19). In the present paper, however, we use the term in its broader sense of a finite verb (with full or defective inflection), which is used in combination with the lexical verb and expresses features such as person, number, and TAM[3].

Focus is the grammatical category that determines which part of the sentence provides new or contrastive information (see further Zuo and Zuo (2001)). In many languages, e.g. in Armenian, instead of (or in addition to) stress, word order can be used to express focus[4], see the example below:

a. *Kat´ołikos-ə* *ut-um* *ēr*
Catholicos-DEF eat-IPFV AUX.PST.3SG

'Catholicos was eating.' (And not doing something else)[5]

b. *Kat´ołikos-n* *ēr* *ut-um*
Catholicos-DEF AUX.PST.3SG EAT-IPFV

'Catholicos was eating.' (It was Catholicos, who was eating.)

2.3. Georgian and Armenian

Eastern Armenian forms some of its tenses by combining certain non-finite forms of the verb with the unstressed *to-be*-auxiliary, which originates from the copula and is inflected for person/number and tense (present/past) (cf. e.g. Comrie (1984); Tamrazian (1991); Kahnemuyipour and Megerdoomian (2017)). While the context-independent citation form of this predicative construction is V AUX, the auxiliary can attach enclitically to any constituent before the main verb in a given context to mark the syntactic focus of the clause. However, it cannot attach to full words following the verb (this was verified by the results of the corpus search, see Section 7.1.).

In Modern Georgian, just as in English, the notions of possibility, necessity and desire are expressed by auxiliary verbs: *unda*[6] 'must', *minda* 'I want', *mč'irdeba* 'I need', *šemiʒlia* 'I can'. Georgian natural sentential word order fluctuates between SOV[7] and SVO (Vogt, 1974)[8] with a preference for OV in shorter sentences (Apronidze, 1986, p. 26). In languages with dominant SOV order, one would expect the auxiliary to follow the main verb (Greenberg, 1963, universal 16). However, a cursory corpus-based investigation (looking at the verbs 'must', 'to want' and 'to be able to' in the GNC[9]) shows that appr. 80% of clauses with an auxiliary show the order AUX V, which corresponds to the citation form of Armenian modal verbs (e.g. *piti gnam* must go.1SG 'I must go').

Thus, the prevalent order is V AUX (where AUX is a form of 'to be') in Armenian and AUX V (where AUX is a modal verb) in Georgian. A further investigation will consider conditions under which word order deviates from these prevalent patterns and the frequency of certain order types. One such condition could be focus, since the element directly before the AUX is expected to have syntactic focus. Furthermore, the influence of different types of focus (besides syntactic focus) could be examined. If both Armenian and Georgian show similar strategies regarding the expression of focus with use of the placement of the auxiliary, syntactic convergence due to language contact could be considered.

In the scope of the present paper, we only conduct a preliminary experiment in order to check the operability of the pipeline.

3. Language Resources

3.1. Eastern Armenian National Corpus

With its 110 million tokens, the Eastern Armenian National Corpus (EANC)[10], contains written texts in different genres (fiction, news, scientific texts, and other non-fiction), transcripts of oral communication, and logs of electronic communication. Nearly all genres are represented as fully as possible (except for the electronic communication and online news). All the texts are morphologically parsed without disambiguation. A tagset used for the corpus was developed specifically for the EANC project.

From a technical perspective, texts are represented in a CoNLL-like format (TSV[11]). The main difference from the traditional CoNLL is the presence of alternative parses: since there is no disambiguation in the EANC corpus, an

[3]Tense, aspect, mood.

[4]Here, we refer only to syntactic focus; Comrie (1984, pp.3-4) distinguishes this from pragmatic and intonational focus.

[5]Vrt'anes P'ap'azyan, Stories. EANC

[6]Although discussion may arise as to whether this word is truly verbal (since it is not inflected), it does fulfill the same function as the other modal verbs.

[7]S(ubject), O(bject), V(erb).

[8]The same uncertainty as to basic SOV-SVO order applies to Armenian, cf. Comrie (1984, p. 4).

[9]Georgian National Corpus, see Section 3.3.

[10]http://eanc.net/EANC/search

[11]Tab-separated values

annotation of each word is repeated for every possible morphological parse. To represent this in CoNLL, the authors output every possible parse as a separate word (on a new line) but with the same word ID. This non-standard format required updating the CoNLL-RDF conversion (see section 4.1.) to correctly handle this design decision.

3.2. Interlinear Glossed Georgian Text in FLEx

Fieldwork Language Explorer (FLEx)[12] is a tool designed for field linguists to create interlinear glossed text and lexicons, and also features some (limited) corpus query functionalities. The user can completely customize its part of speech tagsets, and the glosses of grammatical morphemes can be viewed as further annotation tags. The output is an XML file with the extension .flextext, which contains one annotated text. A collection of short stories by Erlom Akhvlediani (1986) called "Vano & Niko" have been glossed and exported accordingly. This sample consists of approximately 900 sentences and reflects the modern standard Georgian literary language.

3.3. Georgian National Corpus

The Georgian National Corpus (GNC)[13] is developed by researchers at the universities of Frankfurt, Bergen, and Tbilisi, and contains over 227 million tokens. The corpus, which is still under development, contains subcorpora of Old, Middle and Modern Georgian, plus two subcorpora of transcribed recordings of spoken language. Corpora of Megrelian and Svan texts are under construction as well. A large Georgian reference corpus (GRC) is included that contains less thoroughly processed texts from various fictional and non-fictional domains. The Georgian texts (within GNC and GRC) are fully morphologically annotated (lemma forms and morphosyntactic features), and all texts in the GNC subcorpora have comprehensive metadata.

4. Conversion to RDF

In a first step, we convert the source formats to an isomorphic rendering in RDF, which then represents the basis for further normalization.

4.1. CoNLL ⇒ RDF

To facilitate the processing of TSV formats such as the EANC format, the CoNLL format family, or popular infrastructures such as the corpus workbench, the **CoNLL-RDF** package (Chiarcos and Fäth, 2017)[14] uses RDF technology. In this way, it enables the advanced manipulation of annotated corpora (graph rewriting) with SPARQL UPDATE, their quantitative evaluation with SPARQL SELECT, off-the-shelf database support with RDF Triple/Quad Stores, sentence-level stream processing and access with a W3C standardized query language (SPARQL). CoNLL-RDF provides an isomorphic, but shallow reconstruction of CoNLL data structures in RDF:

- Every row — which in standard CoNLL corresponds to a word — is mapped to a *nif:Word* (using the NIF vocabulary, Hellmann et al. (2013)). As mentioned above, the EANC corpus is not disambiguated and therefore, there can appear several lines for one and the same word in the TSV files, each line containing the word with a different possible parse. This problem was solved by joining the different annotations into a triple group containing the same subject (the URI of the word) and predicate (the annotation type), while having several objects — one for each annotation possibility (e.g. `:s1_1 conll:GRAM "cvb conneg", "sbjv pres sg 3", "imp sg 2".`).
- Consecutive words are connected by *nif:nextWord*.
- Rows which are not separated by an empty line are represented as a *nif:Sentence*.
- Consecutive sentences are connected by *nif:nextSentence*.
- The actual annotations in the original CoNLL files are stored in columns. Every column with a user-provided label, say, *WORD, POS*, etc., is rendered as a property in the conll namespace (*conll:WORD, conll:POS*, etc.).

The EANC corpus files and the GNC data are converted to CoNLL-RDF, because the GNC — in addition to its native XML format — is also available in CoNLL-U. An example of the resulting RDF data displayed in the Turtle syntax is given in Fig. 1.

4.2. FLEx ⇒ RDF

For the RDF rendering of the FLEx data, we use the FLEx LLODifier tool,[15] which converts to the so-called FLEx-RDF format. The LLodifier is a collection of tools for converting language resources into an RDF representation (Chiarcos et al., 2017). In comparison to CoNLL, the FLEx data model is complex, as it allows annotations on three levels of granularity: *flex:phrase, flex:word*, and *flex:morph*. These are furthermore organized hierarchically (a *flex:phrase flex:has_word* some *flex:word*, a *flex:word flex:has_morph* some *flex:morph*) as well as sequentially (*flex:next_phrase*, *flex:next_word*, *flex:next_morph*).

5. Harmonization

These different, source-specific RDF renderings of our respective data are now transformed into uniform representations by anchoring them in more general LLOD vocabularies and terminology bases.

To represent linguistic data structures in general, we use POWLA (Chiarcos, 2012), an OWL2/DL reconstruction of the Linguistic Annotation Framework (LAF).[16] From the LAF, POWLA inherits the claim to represent *any* linguistic data structures applicable to textual data.

[12] https://software.sil.org/fieldworks/

[13] http://gnc.gov.ge/gnc/page?page-id=gnc-main-page

[14] implemented in Java and available under Apache 2.0 license, https://github.com/acoli-repo/conll-rdf

[15] https://github.com/acoli-repo/LLODifier/tree/master/flex

[16] https://www.iso.org/standard/37326.html

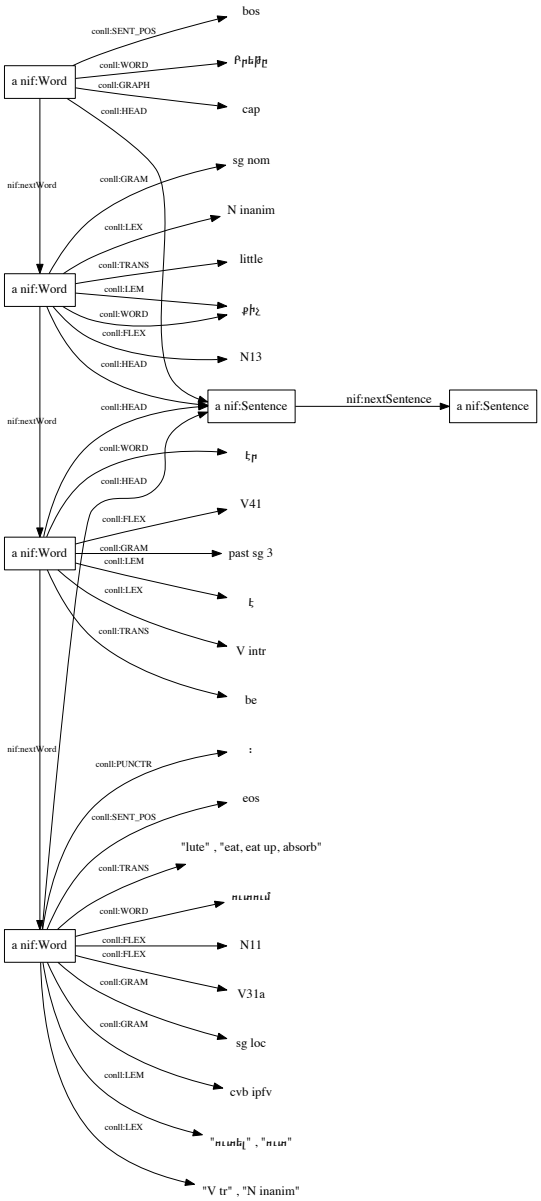

Figure 1: Example of the EANC data converted to CoNLL-RDF

To represent linguistic annotations while guaranteeing interoperability, we apply the Ontologies of Linguistic Annotation (OLiA)[17] which allow us to derive a structured, ontology-based representation from plain tags as used during the annotation.

5.1. POWLA and the LAF

It is generally accepted that any kind of linguistic annotation can be represented by means of directed (acyclic) graphs (Bird and Liberman, 2001; Ide and Suderman, 2007): Aside from the primary data (text), linguistic annotations consist of three principal components, i.e., segments (spans of text, e.g., a phrase), relations between segments

[17] http://purl.org/olia/

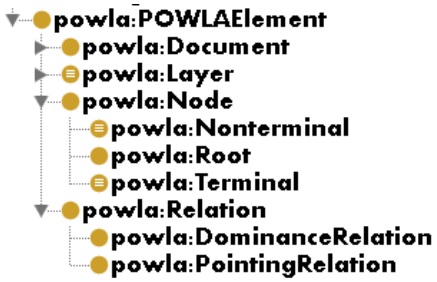

Figure 2: The POWLA data model

(e.g., dominance relation between two phrases) and annotations that describe different types of segments or relations. In graph-theoretical terms, segments can be formalized as nodes, relations as directed edges and annotations as labels attached to nodes and/or edges. These structures can then be connected to the primary data by means of pointers. A number of generic formats have been proposed on the basis of such a mapping from annotations to graphs, most importantly the Linguistic Annotation Framework (LAF) developed by ISO TC37/SC4. Such formats are traditionally serialized as standoff XML, e.g., in the GrAF format, but as these are poorly supported by off-the-shelf technology and highly domain-specific, serializations of this data model in RDF have been developed. Here, we focus on POWLA (Chiarcos, 2012), an OWL/DL serialization of the data model of the PAULA XML format (Dipper, 2005; Chiarcos et al., 2008; Chiarcos et al., 2011), a generic interchange format that originates from early drafts of the Linguistic Annotation Framework, and which is closely related to the later ISO TC37/SC4 format GrAF. PAULA was designed to support the lossless representation of arbitrary kinds of text-oriented linguistic annotation, and in particular the merging of annotations produced by different tools (e.g., multiple independent syntax annotations (Chiarcos, 2010) , or syntax, coreference and discourse structure annotation at the same time, (Chiarcos et al., 2011)). With POWLA, these annotations can also be represented by means of Semantic Web standards.

The POWLA data model, as illustrated here (Fig. 2), is relatively minimalistic. Aside from corpus structure (*powla:Document, powla:Layer*), annotations are grounded in *powla:Nodes* which can be linked by *powla:Relations* (hierarchical dominance relations, or non-hierarchical pointing relations with explicit *hasTarget/hasSource* properties). Hierarchical relations are accompanied by a *powla:hasChild* (resp. *powla:hasParent*) property between the parent and child node, which can also be used without *powla:Relation* for an unlabeled hierarchical relation.

For our use case, POWLA allows us to generalize over both data models (CoNLL-RDF and FLEx-RDF): The mapping of the format-specific *nif/flex* categories into POWLA categories is listed in Tab. 1.

This generalization is done by a SPARQL UPDATE script which loads an ontology providing the rdfs:subClassOf, rdfs:subPropertyOf statements for the FLEx (resp. CoNLL (Fig. 3)) categories. Using this ontology, the update replaces the original CoNLL (FLEx) data structures with

EANC, GNC (via CoNLL-RDF)	**Georgian IGT** (FLEx, via FLEx-RDF)	**POWLA**
nif:Word, nif:Sentence	flex:word, flex:phrase, flex:morph	powla:Node
nif:nextWord, nif:nextSentence	flex:next_word, flex:next_phrase, flex:next_morph	powla:next
conll:HEAD (to nif:Sentence)	flex:has_word, flex:has_phrase, flex:has_morph	powla:hasChild / hasParent

Table 1: Harmonization of corpus formats via POWLA

POWLA data structures (Fig. 4). The actual annotations of these data structures are, however, left in their original namespace, as they are extensible in the original formats/tools. Fig. 5 illustrates an extract of the data resulting from running this SPARQL UPDATE script.

```
...
    <owl:ObjectProperty rdf:about="http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#next
        <rdfs:subPropertyOf rdf:resource="http://purl.org/powla/powla.owl#nextNode"/>
    </owl:ObjectProperty>
...
    <owl:Class rdf:about="http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Sentence">
        <rdfs:subClassOf rdf:resource="http://purl.org/powla/powla.owl#Root"/>
    </owl:Class>
```

Figure 3: Extract of the ontology conllpowla.owl

5.2. Mapping to OLiA

After unifying the data formats by converting to an RDF format and mapping to the POWLA data structure, the values of the annotation must also be harmonized. Therefore, we needed to define mapping rules for *conll:GRAM, conll:LEX, flex:gls,* etc. to a unified annotation model. This is done by employing OLiA (REF), the Ontologies of Linguistic Annotation. It provides:

1. a modular architecture of ontologies for annotation models for different languages,
2. the OLiA reference model and
3. linking models.

The linking models connect the annotation models (1.) to the OLiA reference model with *rdfs:subClassOf* (etc.)

```
# assume that the graph <http://purl.org/powla/> contains a mapping to POWLA
# (cf. conllrdf.owl )
INSERT {
        ?a ?powlaProp ?b
} WHERE {
        ?a ?prop ?b.
        GRAPH <http://purl.org/powla/> {
                ?prop rdfs:subPropertyOf ?powlaProp.
                FILTER(contains(str(?powlaProp), "http://purl.org/powla"))
        }
};

INSERT {
        ?a a ?powlaClass.
} WHERE {
        ?a a ?Class.
        GRAPH <http://purl.org/powla/> {
                ?Class rdfs:subClassOf ?powlaClass.
                FILTER(contains(str(?powlaClass), "http://purl.org/powla"))
        }
};
```

Figure 4: Extract of the SPARQL UPDATE to complement CoNLL-RDF data structures with POWLA data structures

statements. OLiA already provides several annotation models (e.g. for the Universal Dependencies (UD)), but for Georgian FLEx, GNC, and the Armenian EANC data, we had to develop novel annotation models[18].

Since an annotation tag in all the given corpora consists of several features (e.g. "V intr"), we used the *hasTagContaining* property of the OLiA System Ontology[19] to attribute the features to its Named Individual in our annotation models (e.g. `eanc:intr system:hasTagContaining intr^^xsd:string .`). This property, however, is unsuitable for features, whose strings partially coincide with others (e.g. tr for transitive and intr intransitive). To solve this ambiguity, the hasTagMatching property with a regular expression was used instead (e.g. `eanc:tr system:hasTagMatching ^(.*)*tr( .*)*$^^xsd:string.`).

Figure 6 illustrates how the OLiA mapping for a specific tag (in this example marking a cardinal numeral) in the EANC corpus functions by linking the EANC annotation model class (EANC CardinalNumber) to its super class in OLiA (OLiA CardinalNumber).

The implementation of the OLiA mapping is done by a SPARQL UPDATE, similarly to the POWLA mapping. The update inserts unified annotations according to the corresponding annotation model. For the EANC annotation model, the query is shown in Figure 7.

The features used in the GNC (303 in total) have a shallow hierarchy. They are divided into two categories, i.e. Part of Speech, and Grammatical Features, and have been mapped to OLiA as such. Similarly, the tags used in FLEx are divided into PoS (annotated in FLEx as Word Category) and other grammatical features (annotated in FLEx as glosses). Because of the large number of superfluous features, only basic PoS features and their OLiA mapping have been used for the experiment, i.e. Verb, Noun, Modal.

The linking of our annotation models to the OLiA reference model faced certain challenges. On the one hand, the linking requires to find the OLiA category which best generalizes over a language-specific category, and an agreement between specialists of the language needed to be found. On the other hand, the OLiA coverage is by nature incomplete, and when linking a new language which contains concepts not yet covered in OLiA, its extension becomes necessary. This was the case for the Converb, appearing both in the EANC and GNC annotation model. Finally, a class in the annotation model is not always linkable to just one class in OLiA. It can be linked to multiple OLiA classes at once, or there can be several alternative OLiA classes to which one might want to link (e.g. the EANC class Determination/Possession is either a subclass of the OLiA DefiniteArticle or of the OLiA PossessiveDeterminer, but not both.). For the latter case, we use the UNION operator of the Turtle syntax. To retrieve the conjuncts of a UNION in a SPARQL query, one can either just query for the first OLiA

[18]The GNC tagset is currently under revision and will be converted to UD v.2 with some extension (personal communication with Paul Meurer in Nov. 2017). Thus, in the future, our own GNC annotation model will be replaced by the existing annotation model for UD.

[19]`http://purl.org/olia/system.owl`

```
@prefix : <file:///C:/Users/chiarcos/Desktop/corpus/armenian/EANC_sentences_sample///fiction.tsv#> .
@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix powla: <http://purl.org/powla/powla.owl#> .
:s42_0 nif:nextSentence :s43_0;
        powla:nextNode :s43_0 .
:s43_0 a powla:Root , nif:Sentence .
:s43_1 a nif:Word, nif:nextWord :s43_2;
        powla:Terminal; powla:hasParent:s43_0; powla:hasStringValue "Կերուխուս"; powla:nextNode :s43_2;
        conll:WORD "Կերուխուս"; conll:GRAM "sg nom"; conll:HEAD :s43_0; conll:LEM "Կերուխուս"; conll:LEX "N inanim" .
:s43_2 a nif:Word, nif:nextWord :s43_3;
        powla:Terminal; powla:hasParent:s43_0; powla:hasStringValue "են"; powla:nextNode :s43_3;
        conll:WORD "են"; conll:GRAM "pres pl 3"; conll:HEAD :s43_0; conll:LEM "է"; conll:LEX "V intr"; conll:TRANS "be" .
:s43_3 a nif:Word, nif:nextWord :s43_4;
        powla:Terminal; powla:hasParent:s43_0; powla:hasStringValue "պել"; powla:nextNode :s43_4;
        conll:WORD "պել"; conll:GRAM "cvb pfv"; conll:HEAD :s43_0; conll:LEM "ասել"; conll:LEX "V tr"; conll:PUNCTR ","; conll:TRANS "do, make, make up" .
:s43_4 a nif:Word,
        powla:Terminal; powla:hasParent :s43_0; powla:hasStringValue "ուրախացել";
        conll:WORD "ուրախացել";conll:GRAM "cvb pfv"; conll:HEAD :s43_0; conll:LEM "ուրախանալ"; conll:PUNCTL ","; conll:TRANS "be glad/happy, enjoy oneself" .
```

Figure 5: Extract of POWLA annotated CoNLL-RDF data

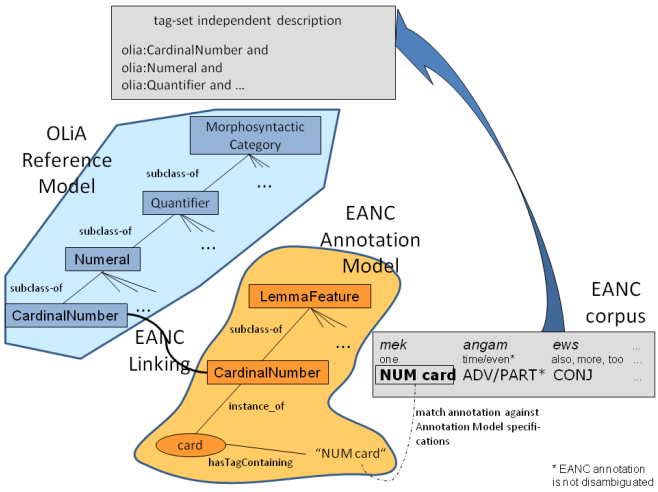

Figure 6: Visualization of the Linking of EANC and OLiA

conjunct (using *rdf:first*), for cases in which a hierarchy is defined stating that the first conjunct is the most probable, or one can extract all the conjuncts (as in the query in Figure 7). Extracting all OLiA conjuncts in order to link an annotation to all of them results, however, in the loss of the information about the conjuncts being mutually exclusive.

6. Experimental Setup

We conducted a case study on word order within auxiliary and main verb constructions. This was first applied to a part of the EANC corpus. In the future, we will replicate it on different Georgian corpora, i.e., the Georgian National Corpus (GNC) and interlinear glossed data (see Section 3.).

6.1. Pipeline

As described above, we first convert the corpora to shallow RDF-representations (CoNLL-RDF / FLEx-RDF). Then, we harmonize the data structures by transforming them to POWLA (Section 5.). This is followed by bringing the different annotation schemes of each of the corpora together through the concept linking with the OLiA tagset (Section 5.2.).

Through the harmonization of the data formats and the linking of the language specific annotations to OLiA, we are able to combine all our resources. The resulting RDF data for each of our corpora can then be queried in a unified manner. We can also add triples containing intermediate query results in order to execute advanced queries faster by using these intermediate results. In our research, we added triples containing the information about a word being an auxiliary or a main verb (according to the language specific definitions) and in a following query, we use this information to analyze the word order. The full pipeline for converting, unifying and getting experimental data is illustrated in Fig. 8.

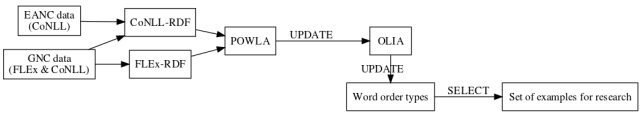

Figure 8: Pipeline of the experiment

The enriched annotation can be used for further qualitative linguistic analysis. We extract candidate sentences with a SPARQL SELECT and then study the distribution of different auxiliary / main verb ordering types manually as a preparation for a future analysis of the word order / focus implications (see Section 2.).

6.2. Scope of the Experiment

We restrict this experiment to the extraction and classification of structurally / morphologically unambiguous cases. In a future research, however, we plan to extend it to more complex sentence structures. Conceptual difficulties of our experiment are the comparability of the types of auxiliaries in the two languages and common complications in the annotation of the corpora for both languages, such as the absence of syntactic annotation and non-disambiguation on the morphological level. The problem of the (natural) shortage in the OLiA-terminology (i.e. absence of the concept converb) was solved by the extension described in Section 5.2.. The linguistic outcomes of the research are preliminary and serve only to exploring a hypothesis. A full-fledged linguistic investigation requires additional annotation efforts.

In the following, we illustrate these steps for Armenian. The pipeline scripts will be published via our GitHub repository[20] under an open license.

6.3. Filtering Clauses

We only considered sentences containing no further tokens tagged as verb beside the auxiliary and the main verb. There

[20]https://github.com/acoli-repo

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

INSERT {
        ?x a ?super
} WHERE {
        ?x a ?eanc_annot .
        GRAPH <http://example.org/eanc.owl> {
                ?eanc_annot a owl:Class .
                FILTER(contains(str(?eanc_annot), "http://purl.org/olia/eanc.owl"))
        }
        GRAPH <http://example.org/eanc_link.ttl> {
                ?c rdfs:subClassOf+/(rdfs:subClassOf*/(owl:unionOf/(rdf:first|(rdf:rest*/rdf:first)))/rdfs:subClassOf*)? ?super .
                FILTER(contains(str(?super),'http://purl.org/olia/olia.owl'))
        }
        FILTER(?c = ?eanc_annot)
}
```

Figure 7: SPARQL UPDATE for OLiA mapping (EANC)

```
PREFIX conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX olia:  <http://purl.org/olia/olia.owl#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

INSERT {
        ?a rdfs:comment "auxiliary" .
        ?b rdfs:comment "main verb" .
} WHERE {

        #get auxiliary according to definition
        ?a powla:hasParent ?s .
        ?a conll:LEM  't'; a olia:Verb .

        { ?b powla:nextNode+ ?a . } UNION { ?a powla:nextNode+ ?b . }

        #get main verb according to definition
        ?b a ?e1 , ?e2 .
        FILTER(?e1 = olia:Gerund)
        FILTER(?e2 = olia:ImperfectiveAspect || ?e2 = olia:PerfectiveAspect

        #get only simple sentences
        FILTER NOT EXISTS {
                ?another_verb powla:hasParent ?s; ?another_verb a olia:Verb
                FILTER (?another_verb != ?a && ?another_verb != ?b)
        }
        MINUS {
                ?q powla:hasParent ?s;  powla:hasStringValue ?question .
                FILTER (contains(str(?question),"^"))
        }
        MINUS {
                ?neg powla:hasParent ?s; conll:LEM  't'; a olia:Negation.
        }
}
```

Figure 9: Example SPARQL update for auxiliary and main verb annotation for the EANC data

```
PREFIX conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

INSERT {
        ?s rdfs:comment ?wo_comment .
} WHERE {
        ?a powla:hasParent ?s ; ?a rdfs:comment "auxiliary".

        {
        #main verb directly after auxiliary
        ?a powla:nextNode ?b .
        BIND("auxiliary directly before main verb" as ?wo_comment)
        } UNION {
        #main verb NOT directly after auxiliary
        ?a powla:nextNode/powla:nextNode+ ?b .
        BIND("auxiliary NOT directly before main verb" as ?wo_comment)
        } UNION {
        #main verb directly before auxiliary
        ?b powla:nextNode ?a .
        BIND("auxiliary directly after main verb" as ?wo_comment)
        } UNION {
        #main verb NOT directly before auxiliary
        ?b powla:nextNode/powla:nextNode+ ?a .
        BIND("auxiliary NOT directly after main verb" as ?wo_comment)
        }

        ?b rdfs:comment "main verb" .
}
```

Figure 10: SPARQL UPDATE for word order annotation

are some language specific filters to be taken into consideration in order to extract correct examples (see Section 2.3.); e.g. for the auxiliary to be in Armenian, we can only consider sentences in which this auxiliary (recognizable by its lemma (*conll:LEM*)) is combined with a main verb in certain tenses, in which it is not negated etc. A simplified SPARQL UPDATE to mark auxiliary and main verb with a *rdfs:comment* according to these filters is given in Fig. 9.

6.4. Classifying Clauses

Having added the *rdfs:comment* triples to the auxiliary and main verbs language-specifically for the EANC, GNC and FLEx data, the classification of the sentences with respect to the word order of these verbs can be done language-independently by the SPARQL UPDATE script shown in Figure 10: The word order information is also added by inserting *rdfs:comment* triples.

After annotating the selected sentences with their word order features (auxiliary directly/not directly before/after main verb) as a *rdfs:comment*, we export them to a CSV file (using a SPARQL SELECT query which filters out all sentences not annotated with a word order feature) containing the sentences themselves, their genre and their word order type including the position of the auxiliary and main verb. In such a restricted table format, a qualitative analysis of the relation between word order type and focus marking is facilitated and can be done more efficiently than in the underlying RDF format containing triples which are only relevant for the comparability/harmonization of the data and constitute redundant information in the perspective of a qualitative analyser.

7. Discussion

So far, we described the general setup of our approach, its technological components, and the data sets. While a full-fledged linguistic interpretation of our findings is beyond the scope of this paper and will be forthcoming , an evaluation in quantitative terms has been conducted.

7.1. Quantitative Evaluation

Applying the limitation and filters mentioned in Section 6.2. and 6.3., we get 20 043 classified sentences corresponding to 8.13% of the entire EANC subcorpus on hand (246,678 sentences in total). The manual evaluation consisted in examining a subset of classified sentences in order to determine the ratio of false positives and what technical and/or filter shortages caused their occurrence, if any. The distribution of the word order types among the classified sentences as well as the results of the manual evaluation are shown in Table 2.

The occurrence of false positives is mostly due to the non-disambiguated annotation of the EANC. This is especially

Word Order Type	Number of sentences	Manually evaluated sentences	Precision %
AUX V	4,993	303	95.38
V AUX	14,494	300	99.67
AUX * V	540	152	36.18
V * AUX	16	16	0
total	20,043	771	83.40

Table 2: The distribution of word order types in predicative constructions with a *to-be*-AUX in EANC and the results of manual evaluation. The * means, that there is at least one element between V(erb) and AUX(iliary)

the case with the AUX * V word order type[21]. The number of false positives is almost evenly distributed in all of the three genres (fiction, non-fiction, press)[22]. We do not calculate the recall, as it would require to manually search the remaining 226 635 sentences (91.87%) of the subcorpus. However, an analysis of 250 non-classified sentences with 23 false negatives shows that the latter are likewise in most cases due to the non-disambiguated annotation of the EANC. Less than the half of the false negatives also include further non-finite verbs. To exclude these, further filter restrictions must be considered to refine the search later on.

7.2. Conclusion and Outlook

We demonstrated how to employ LLOD formalisms to develop extraction pipelines for features and examples from diverse and heterogeneous corpora and collections of interlinear glossed text. Originally available in different formats, RDF, SPARQL and LLOD vocabularies facilitate unified access, enrichment and exploitation of such data.

After conversion from the original formats to an isomorphic, and semantically shallow RDF representation, SPARQL UPDATE can be applied to conveniently transform the original data to a common data model (here, POWLA). Similarly, SPARQL UPDATE allows to load external ontologies, and with the annotation models for EANC, IGT and GNC that we contribute to OLiA, we can follow their links with SPARQL property paths and render linguistic annotations in terms of ontological concepts.

As a result, extraction and transformation pipelines can be developed for this data, and to the extent that annotations are comparable both in terms of their hierarchical organization and in terms of their linguistic expressiveness, extraction (or transformation) scripts can be applied to other corpora in other languages.

Even after POWLA conversion, however, interpreting the original data structures is not without complications: The hierarchical nesting of *powla:Nodes* in different corpora (e.g. on the level of morphs in FLEx, but on the level of words in CoNLL-RDF) poses difficulties in following *powla:next* immediately. However, as long as we are dealing with trees, and as long as siblings (and siblings only) are always connected by a *powla:next* property, this generalized precedence operator between two variables `?x` and `?y` can be defined by the following SPARQL property path:

```
?x powla:hasParent*|powla:next+|powla:hasChild* ?y.
```

Immediate adjacency is slightly more complicated, and can be implemented by requiring that no intermediate variables exist:

```
MINUS {
?x powla:hasParent*|powla:next+|powla:hasChild* ?t.
?t powla:hasParent*|powla:next+|powla:hasChild* ?y.
}
```

As these property paths can be time-consuming, we can use SPARQL UPDATE to add a triple, say `?x my:next ?y`, for all immediately adjacent *powla:Node*s, and then use this as a shorthand in subsequent queries. This is, indeed, a key advantage of RDF, which allows to use SPARQL UPDATE to pre-compile costly expressions, thereby speeding up the eventual search process.

The impact of this functionality can only be assessed in comparison with state-of-the-art approaches in corpus linguistics: In order to generalize over different source formats, standoff XML formats (Ide and Suderman, 2007) are still considered the state of the art, but their support with off-the-shelf database technology and APIs is known to be limited (Eckart, 2008). Accordingly, corpus management systems with standoff functionality convert standoff XML to an internal, relational database scheme (Zeldes et al., 2009). However, this means that search in such systems and the retrieval of examples is constrained by a static data model and by pre-defined optimizations for a particular type of query (or lack thereof). Using RDF and SPARQL, shorthands can be introduced during the search at any point in time (if the database permits).

Our approach has been successfully implemented and described here for the study of syntactic convergence phenomena in genetically unrelated, but neighboring languages from the Caucasus area, Armenian and Georgian. It is, however, not limited to this task, and can be applied to other linguistic research questions, as well.

While this demonstrates the functionality and the technological appeal of our approach, it must be noted that SPARQL and RDF are not *a priori* linguist-friendly formalisms and technologies. One goal of our project is to facilitate the accessibility and usability of LOD technology for linguists. By demonstrating that these are viable technologies for linguistic problems, and that they allow to overcome technical barriers that currently limit the joint evaluation of available linguistic data sets in an unprecedented way, we can now motivate increased efforts in developing LOD-based infrastructures for linguistic research questions.

[21]The ambiguity is due to the fact that imperfective has the same suffix as the locative case, and infinitive and perfective of some verbs concur in form.

[22] Fiction: 32/252; non-fiction: 51/255; press: 45/256.

8. Bibliographical References

Apronidze, S. (1986). *sitqvatganlageba axal kartulši. marṭivi ċinadadeba (Modern Georgian word order. The simple clause)*. Tbilisi Academic Press.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, et al., editors, *Language, Data, and Knowledge. LDK 2017. Lecture Notes in Computer Science, vol 10318*, pages 74–88, Cham, Switzerland, June. Springer.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., and Stede, M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *TAL (Traitement automatique des langues)*, 49(2):217–246.

Chiarcos, C., Ritz, J., and Stede, M. (2011). Querying and visualizing coreference annotation in multi-layer corpora. In Iris Hendrickx, et al., editors, *Proceedings of the Eigth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 80–92, Faro, Portugal, October. Edições Colibri.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards Open Data for Linguistics: Linguistic Linked Data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing*, pages 7–25. Springer, Berlin, Heidelberg, Germany.

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C., Wichers Schreur, J., and Nevskaya, I. (2017). LLODifying Linguistic Glosses. In *Language, Data, and Knowledge. LDK 2017. Lecture Notes in Computer Science, vol 10318*, pages 89–103, Cham, Switzerland, June. Springer.

Chiarcos, C. (2010). Towards Robust Multi-Tool Tagging. An OWL/DL-Based Approach. In Jan Hajič, et al., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 659–670, Uppsala, Sweden, July. Association for Computational Linguistics.

Chiarcos, C. (2012). POWLA: Modeling Linguistic Corpora in OWL/DL. In Elena Simperl, et al., editors, *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 225–239. Springer, Berlin, Heidelberg, Germany.

Comrie, B. (1984). Some formal properties of focus in Modern Eastern Armenian. *Annual of Armenian Linguistics*, 5:1–21.

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In Rainer Eckstein et al., editors, *Berliner XML Tage 2005*, pages 39–50, Berlin, Germany, September. Humboldt-Universität zu Berlin.

Eckart, R. (2008). Choosing an XML database for linguistically annotated corpora. *Sprache und Datenverarbeitung, International Journal for Language Data Processing (SDV)*, 32(1):7–22.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, USA.

Harris, A. C. (1981). *Georgian syntax: a study in relational grammar*. Cambridge University Press.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP Using Linked Data. In David Hutchison, et al., editors, *The Semantic Web – ISWC 2013*, volume 8219 of *Lecture Notes in Computer Science*, pages 98–113. Springer, Berlin, Heidelberg, Germany.

Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *1st Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, Prague, Czech Republic, June.

Kahnemuyipour, A. and Megerdoomian, K. (2017). On the positional distribution of an armenian auxiliary: Second-position clisis, focus, and phases. *Syntax*, 20:77–97.

Ramat, P. (1987). Introductory paper. In Martin Harris et al., editors, *Historical development of auxiliaries*, pages 3–19. Mouton de Gruyter, Berlin, New York, Amsterdam.

Tamrazian, A. (1991). Focus and wh-movement in Armenian. *University College London Working Papers in Linguistics*, 3:101–121.

Vogt, H. (1974). L'ordre des mots en géorgien moderne. In Even Hovdhaugen et al., editors, *Linguistique caucasienne et arménienne*. Norwegian University Press, Oslo.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In Michaela Mahlberg, et al., editors, *Proceedings of the Corpus Linguistics Conference*, pages 20–23, Liverpool, UK, July.

Zuo, Y. and Zuo, W. (2001). *The computing of discourse focus*. Lincom Europa, Munich, Germany.