Universal Morphologies for the Caucasus region

Christian Chiarcos Kathrin Donandt Maxim Ionov
Monika Rind-Pawlowski Hasmik Sargsian Jesse Wichers Schreur
Frank Abromeit Christian Fäth
{chiarcos|donandt|ionov|abromeit|faeth}@informatik.uni-frankfurt.de
{sargsian|wichersschreur}@em.uni-frankfurt.de
{rind-pawlowski}@lingua.uni-frankfurt.de
Goethe University Frankfurt, Germany

Abstract

The Caucasus region is famed for its rich and diverse arrays of languages and language families, often challenging European-centered views established in traditional linguistics. In this paper, we describe ongoing efforts to improve the coverage of Universal Morphologies for languages of the Caucasus region. The Universal Morphologies (UniMorph) are a recent community project aiming to complement the Universal Dependencies which focus on morphosyntax and syntax. We describe the development of UniMorph resources for Nakh-Daghestanian and Kartvelian languages as a well as for Classical Armenian, we discuss challenges that the complex morphology of these and related languages poses to the current design of UniMorph, and suggest possibilities to improve the applicability of UniMorph for languages of the Caucasus region in particular and for low resource languages in general. We also criticize the UniMorph TSV format for its limited expressiveness, and suggest to complement the existing UniMorph workflow with support for additional source formats on grounds of Linked Open Data technology.

Keywords: morphology, Caucasus, UniMorph

1. Background

The Universal Morphology project (Sylak-Glassman et al., 2015b, UniMorph)[1] is a recent community effort aiming to complement the Universal Dependencies (Nivre and others, 2015, UD),[2] which focus on syntax, with coverage of morphology. We describe the development of UniMorph resources for languages of the Caucasus region, known for its rich and diverse arrays of languages and language families, and often posing challenges to European-centered views established in traditional linguistics. In particular, we focus on Nakh-Daghestanian (North-East Caucasian) and Kartvelian (South Caucasian) languages, as well as on Classical Armenian, and discuss challenges that these and related languages pose to the current design of UniMorph.

A practical challenge for linguists working with dictionary data consists of linking it with text data. Corpus-based research thus requires computational models of the morphology of the languages under consideration, i.e., lemmatization, at least. But also for low-resource languages (for which few or small amounts of corpus data exist or have to be collected), an explicit treatment of morphology is necessary for the study of language contact, especially if morphologically rich languages are involved (as in the Caucasus area): Neither inherited words nor loan words are transferred between language(stage)s in their base form only. Accordingly, the computational handling of complex morphological processes and features are important for grasping interrelations of Caucasian languages.

The over 100 languages spoken in the Caucasus are grouped into several language families, out of which three are indigeneous, i.e., Caucasian in a strict sense: Kartvelian or South Caucasian, Abkhazo-Adyghean or (North-)West Caucasian and Nakh-Daghestanian or (North-)East Caucasian. A fourth language family with roots in the Caucasus, Hurro-Urartian, is known only from epigraphic records and assumed to be extinct for more than 2000 years.

With respect to morphosyntax, certain typological traits are frequently encountered in Caucasian languages: (Klimov, 1994)[3]: (1) use of agglutination, with a varying degree of inflective elements, (2) verbocentric sentence structure and complex verbal morphology, often including agreement with multiple syntactic arguments, (3) features of ergative, where the subject argument of intransitive verbs receives the same morphological case as the object of transitive verbs (absolutive case), whereas the transitive subject receives ergative case[4], and (4) in Nakh-Daghestanian languages: rich case systems, with up to more than 40 morphological cases. In addition, all living languages in the Caucasus are low-resource (except for Georgian and Armenian which have considerable amounts of written literature), and many exhibit traces of intense language contact with Iranian, Armenian, Georgian, Turkic, Arabic and/or Russian (reflecting shifting patterns of political dominance in the last 2,500 years).

2. Universal and language-specific morphology

Following the success of the Universal Dependencies as a growing community project, a similar effort for the de-

[1] http://unimorph.github.io/

[2] http://universaldependencies.org/

[3] These characteristics do not apply to Armenian, which is an Indo-European language, albeit 'as Caucasian as an Indo-European language could possibly become' (Gippert, p.c., May 2017).

[4] In addition, active-inverse structures can be found in several Caucasus languages, as manifested, for example, in the Kartvelian 'narrative' case (which is, however, often referred to as 'ergative' in Western linguistics).

velopment of cross-linguistic features for inflectional morphology has been initiated: Universal Morphology. Both projects aim to develop features and categories which are *cross-linguistically applicable* (not necessarily universal in the sense of any notion of 'universal grammar'). As such, the UniMorph annotation schema "allow[s] any given overt, affixal (non-root) inflectional morpheme in any language to be given a precise, language-independent definition ... [by means] of a set of features that represent semantic "atoms" that are never decomposed into more finely differentiated meanings in any natural language" (Sylak-Glassman et al., 2015b, p.674).

2.1. UniMorph inventories

The UniMorph data format is a list of tab-separated values for one word per line, with columns for the word form, the lemma and morphological features; it is thus roughly comparable to the CoNLL format as previously used for, e.g., syntactically annotated corpora of Classical Armenian (Haug and Jøhndal, 2008).[5] The primary data structure of UniMorph is an unordered set of semicolon-separated, unqualified features. Figure 1 shows an example of a conventional gloss of the Megrelian word *kešerxvaduk* 'I will meet you' together with its UniMorph representation.

UniMorph resources are rarely original resources, but rather extracted from existing material,[6] such as Wiktionary (Kirov et al., 2016, first-generation UniMorph inventories) and other dictionaries, bootstrapped from morpheme inventories or corpora (as described here), or generated by rule-based morphologies. However, this conversion-based approach means that the segmentation and annotation principles of the underlying resource tend to be preserved.

In general, UniMorph follows a word-based approach to morphology where inflected forms are organized in paradigms, but their internal structure left unanalyzed. In language documentation, however, a morpheme-based approach prevails, i.e., words are segmented into morphemes which are annotated with the linguistic features that they encode. This can lead to vastly different analyses: In morpheme-based annotation, a number of language-specific features are inflectional morphemes that *contribute* to the indication of morphosemantic features rather than to unambiguously *indicate* them. As such, two Megrelian morphemes in Fig. 4b conspire with other TAM markers to indicate tense, aspect and mood (resp., valency). Which morphosemantic (UniMorph) category these morphemes resolve into, remains, however, unspecified as it cannot be automatically deducted from the original resource. We thus describe their function by means of language-specific labels, here LGSPEC6 and LGSPEC7.

[5] https://proiel.github.io/

[6] The primary reason is that UniMorph morpheme inventories are actually rather uninformative, as the format does not permit to provide translations, examples or metadata, e.g., regarding the source of a particular form. Any serious effort to create morpheme inventories in the context of language documentation or philology thus requires an extended format, from which UniMorph TSV files are then to be extracted.

ma	**si**	**kešerxvaduk**					
ma	si	ke-	še-	r-	xvad	-u	-k
1SG	2SG	AFF	PV	O2	meet	TM	S1/2SG

```
xvad kešerxvaduk AFF;V;LGSPEC4;ARGDA2S;
ARGNO1S;LGSPEC6
```

Figure 1: Megrelian *(ma si) kešerxvaduk* 'I will meet you' as conventional interlinear glossed text (above) and in UniMorph LEMMA - FORM - FEATS representation (below)

2.2. Caucasian languages in UniMorph

Already during the design of the UniMorph guidelines (Sylak-Glassman, 2016), Nakh-Dagestanian languages have been taken into consideration for some phenomena, e.g., with respect to the 'universal' gender features NAKH1, ..., NAKH8 for Nakh-Dagestanian noun classes (Sylak-Glassman, 2016, p.27). Selected features of Abkhazo-Adyghean (on argument marking, p.12-13; on interrogativity, p.29), and Kartvelian (on evidentiality, p.25) have been mentioned, too. Beyond this, languages from the Caucasus area are not discussed in relation to the UniMorph schema and the UniMorph repositories comprise datasets for only Modern Georgian and Modern East Armenian. The datasets provided as result of our efforts thus constitute a major increase in coverage of languages of the Caucasus area. We created morphologically annotated datasets in the UniMorph data format for Megrelian (Kartvelian), Khinalug (Nakh-Daghestanian) and Classical Armenian (Indo-European). Additional data on Batsbi (Nakh-Daghestanian) is in preparation.

2.3. Language specific features

In addition to universal features, UniMorph conventions permit language-specific features to be represented by LGSPEC, followed by a numerical index. Although a separate file that defines those markers can be provided, limiting LGSPEC markers to numerical labels impedes the readability of this data, as the Megrelian example in Fig. 1 illustrates.

For languages with a greater number of language-specific features, this convention for the nomenclature of language-specific features may become problematic, as they are likely to be confused and errors in LGSPEC assignment cannot be easily spotted. A more transparent solution would thus be to allow extended LGSPEC labels with human-readable and established abbreviations as those used in conventional glossing. While difficulties in the choice of labels can be resolved relatively easily,[7] a more severe issue exists with respect to another characteristic of UniMorph, the requirement that features are both unordered and unqualified. Furthermore, when analyzing specific languages, descriptive linguists try to use terminology that fits the grammati-

[7] Resolving such difficulties requires a consensus in the UniMorph community to improve their labeling system or, alternatively, to develop and to document (language-specific) conventions how to deal with conflicting terminologies.

cal phenomenon under question best. Very often, the names and labels that are used are not particularly well suited for cross-linguistic comparison. For example the Kartvelian so-called 'thematic marker' (TM in Fig. 1), is a conventional label given to a morpheme that shows up in certain tense/aspect stems and not in others, but cannot be linked to a specific grammatical function.[8] Consequently, linguists would feel the need to use UniMorph's LGSPEC feature abundantly, to a point where cross-linguistical comparison (be it computational or not) would be impossible. Indeed, it is understood among many linguists that using a single label for two similar grammatical categories in two different languages can be misleading (Haspelmath, 2007). For example, the dative case in Kartvelian can function as the subject, direct object, and indirect object of a clause, which is fundamentally different from its prototypical function (in, say, Latin). The label 'Ergative case' is used in Megrelian to describe a morpheme that not only marks transitive objects, but also intransitive ones. Again, this use contrasts with its prototypical use.

If we follow this line of thought *in absurdum*, all grammatical features would be best translated into UniMorph's LGSPEC, making the procedure pointless. A solution might be found in Linked Data technology (see Section 5.1.). If grammatical features, once translated into UniMorph terminology (in which language-specific details of certain grammatical categories inevitably get lost), would still retain their link to their corresponding language-specific links in the original resource, the procedure would be less lossy, and when necessary, the original data would be easily retraceable.

3. On nominal inflection

3.1. Complex patterns of case marking

In nominal morphology, several instances of the same feature can be overtly realized and need to be distinguished. We discuss this for the double marking of case, which may arise, for example, for languages that provide morphological marking for the inherent case of a noun (reflecting the syntactic status of the noun), and the head case (reflecting the syntactic status of its head). At the moment, instances of such double-coding in nominal morphology are not covered by UniMorph.

Suffixaufnahme: The phenomenon of *Suffixaufnahme* was originally described for Old Georgian and Hurro-Urartian, but has also been documented for **Megrelian** (Boeder, 1995, p.194):[9]

(1) *gi-∅-a-ntχū-d-esə* *k'ata-sə* *te*
PV-O3-LOC-fall-IPF-S3PL.PST people-DAT this
χenc'əpe-ši *ɣəmala-ši-sə*
king-GEN dominion-GEN-DAT

'They attacked the people of this king's dominion.'

Here, *dominion* is not only marked for its inherent case (genitive), but also expresses the (dative) case of its head (*people*). With features regarded as a set (and thus, order-insensitive) of unqualified features, as defined in UniMorph, this information can only be preserved if inherent (genitive) case and head (dative) case are distinguished by different features.

Case attraction: A similar differentiation between inherent case and head case can be found in **Classical Armenian**, although without double-coding (Hübschmann, 1906, p.478-480):

(2) a. *i* *knoǰ-ê* *t'agawor-i-n*
by wife-ABL.SG king-GEN.SG-DEF

b. *i* *knoǰ-ê* *t'agawor-ē-n*
by wife-ABL.SG king-ABL.SG-DEF

'by the wife of the king'

Although Classical Armenian does not mark inherent and head case simultaneously, the regular (inherent) genitive case marking (2a) can be replaced by the morphological case of its head, especially for ablative (2b) or instrumental (Plank, 2014, p.20-21). For the annotation of corpus data, it would be important to distinguish inherent and head case, as they have an impact on syntactic parsing. Beyond the future alignment with the Universal Dependencies, this does not directly concern UniMorph, because the overtly realized case uses the same morphemes for, say, ablative, regardless of whether it indicates inherent or head case. Some languages do, however, provide separate sets of morphemes for both functions, and in these circumstances, it would be important to distinguish inherent case morphology from agreement-based case morphology.

Case combination: Another source of multiple case marking is case combination as found in **Khinalug**: Two cases suffixes can be combined in order to complement their functions. For example, when the ablative in *-(i)lli* attaches to nominal stems directly, it expresses the general ablative meaning, e.g. *mɯda-lli* mountain-ABL 'from the mountain'. However, it can also attach to three other cases. The case in -χ expresses both apudessive and approximative. When combining with the ablative, this leads to the expression of a movement 'away from near sth.', e.g.:

(3) a. *t'u* *quš*
REMT.REF1.ABV bird.ABS
mɯda-χ *učmuškui-'o-mä.*
mountain-APUD/APPRX fly.PRS-ABV-DECL

'The bird up there is flying towards the mountain.'

b. *t'u* *quš*
REMT.REF1.ABV bird.ABS
mɯda-χ-illi
mountain-APUD/APPRX-ABL
učmuškui-'o-mä.
fly.PRS-ABV-DECL

'The bird up there is flying away from near that mountain.'

[8] The label 'present/future stem formant' is misleading since it also shows up in the imperfective past.

[9] This example is likely to be a loan translation from Old Georgian.

The situation is more complex with the possessive-locative in *-š*. Among several other functions, this case attaches to recipient of the verb 'give', when the item is given away for temporary possession only (otherwise the recipient is dative-marked).[10] The ablative-marked form in *-š-illi* marks the former possessor with verbs of the meaning 'take, buy': It combines the meaning of temporary possession (*-š*) and a movement away from the possessor (*-(i)lli*).

(4) a. *Äħmäd-iš* *vaz* *läk'-šä-mä.*
Ahmad-POSLOC knife.ABS give-PST-DECL
'(S/he) gave a knife to Ahmad.'

b. *Äħmäd-iš-illi* *vaz*
Ahmad-POSLOC-ABL knife.ABS
t^{h}enʒuqχ-šä-mä.
buy-PST-DECL
'(S/he) bought a knife from Ahmad.'

Contact-induced double marking: While *-š-illi* in the examples above can be explained by the composition of morphological functions, younger speakers of **Khinalug** prefer to attach *-(i)lli* to *-š* whenever the construction is expressed by an ablative in Azerbaijani – apparently due to the influence of the dominant language. This can be observed for partitive, material indications, and topics of a conversation (even though the construction without ablative is still considered grammatically correct):

(5) *dä* *vaz* *ura-š(-illi)*
ABS.PROX/PHOR knife.ABS iron-POSLOC-ABL
k^{h}ui-qo-mə.
make.PST-FH/BEL-DECL
'This knife was made from iron.'

A similar combination of ablative and comparative case does not seem to be supported by existing case combinations in Khinalug, but triggered by the Azerbaijani use of the ablative for the marking of the object of comparison, since a functional difference between the comparative case marker *-q'* and the comparative-ablative case marking *-q'-illi* cannot be detected at all.

(6) *pši* *hilam-iq'(-illi)* *čiχi*
horse.ABS donkey-COMP-ABL big
qo-mä.
COP.FH/BEL-DECL
'A horse is bigger than a donkey. '

Double case marking in pronominal morphology: Patterns of multiple case marking can also be found in the **inflection of pronouns** in many languages, including well-studied European languages. The German possessive demonstrative pronoun *deren* (roughly, 'their') carries double case marking: In *mit deren Männern* 'with their men', the demonstrative expresses agreement with the head noun *Männer* (DAT;PL) as expected from German adjective (cf. *mit vielen Männern* 'with many men') and article (*mit den Männern* 'with the men') inflection. At the same time, however, *deren* is an extension of an inflected demonstrative, itself, namely from *der* (GEN;SG;F or GEN;PL) – as can be seen from its masculine/neuter counterpart *dessen* (roughly 'his' or 'its', from *des*, GEN;SG;M or GEN;SG;N). Here, the demonstrative carries double inflection: the inherent case, person and number of its antecedent, plus case, person and number of its syntactic head.

Case stacking: While in European languages, this phenomenon is restricted to function words, and therefore of limited relevance to UniMorph, multiple case marking of *full* nouns has also been documented outside the Caucasus, e.g., in Sumerian case stacking:[11]

(7) *ama* *^{D}ba-u$_2$* *e$_2$-tar-sir$_2$-sir$_2$-ta* ...
[mother Bau [E-tarsirsir]$_{ABL}$]$_{ERG}$ ...
'Mother Bau from E-tarsirsir (granted well-being to Gudea).'

Here, the last noun of a noun phrase carries the agreement information of all embedded nouns; the ergative case of *mother Bau* is thus expressed on the last noun of the noun phrase, which itself stands in ablative case, thus resulting in double case marking on *E-tarsirsir*.
Multiple case marking is not limited to two cases. Hurrian *eġli=ve=NE=ve=NA=až=(v)a* 'of the saviour', lit. 'of the one of the salvation', exhibits double genitive, augmented with dative agreement with its head (Wegner, 1995, p.144-145), and, similarly, Sumerian *bi$_3$-lu$_5$-da ud-bi-ta* 'customs of former times' is literally 'ritual of from-the-day', i.e., 'day.ABL.GEN.ABS'.[12]
Such examples can be taken as instances of **recursive inflectional morphology** (Kracht, 2003), which the current design of UniMorph as a flat, unordered set of unqualified features, however, cannot express.

Locatives Multiple case marking is discussed in the UniMorph schema only with reference to locatives: For these, Sylak-Glassman (2016, p.18) follows the analysis of Radkevich (2010, p.5), who suggests the following universal template for the arrangement of local cases:

Noun.Lemma-Stem.Extender-Place-Distal-Motion-Aspect

Non-locative cases are thus expected to occur as stem extenders, and he even mentions Nakh-Dagestanian languages as an example for the use of ergative marking at this point. This description, however, leads to the impression that multiple case marking only occurs in a constellation

[10] Other functions include to mark the addressee of the verb *li* 'say', the topic of a conversation with other *verba dicendi*, it may also function as a partitive and mark the material something is made of, and it marks the subject of abilitative predicates.

[11] http://oracc.museum.upenn.edu/etcsri/Q001547, iii.2-5

[12] http://oracc.museum.upenn.edu/etcsri/Q001124, vii.26-27

where one inherent, non-locative case is applied together with some locative marking. If that would be true, multiple cases could always be resolved unambiguously, in that any non-locative case is by definition the inherent case. However, the examples given above involve several non-locative cases or even a reversal of this pattern, so that an explicit mechanism to express multiple case marking is required.

3.2. A ranking mechanism

As a conservative extension of UniMorph, we suggest to introduce numerical indices to 'non-default' agreement features in nominal morphology. Default agreement would be the inherent case of a noun or pronoun: In a possessive construction like 'people-DAT of this king's-GEN dominion-GEN-DAT' (see the Megrelian example above), the *dominion* would thus receive GEN as a mark of its inherent case. Case features that indicate the agreement of the immediate head would then receive index 1, thus marking *dominion* with GEN;DAT-1 because of its dative agreement with *people*. To express an agreement with the head of the head, the index increases accordingly: Hurrian *eġli=ve=NE=ve=NA=až=(v)a* would thus be glossed as N;GEN;GEN-1;DAT-2, etc.

In preparation for the syntactic annotation of corpora of Classical Armenian and the example given above, we can now also distinguish between *t'agawor-i-n* (N;GEN;SG;DEF) and *t'agawor-ē-n* (N;ABL-1;SG-1;DEF), or between the latter and the same word expressing inherent case (N;ABL;SG;DEF).

In order to gloss the Khinalug double case marked forms appropriately, we suggest to indicate the first case suffix as usual, and add a number to any further following case suffix starting with 1: The tokens in question (in examples (3b), (4b), and (6)) would thus be glossed as follows: *mɯda-χ-illi* N;APUD/APPRX;**ABL-1**, *Äħmädiš-illi* PROPN;LGSPEC.POSLOC;**ABL-1**, and *hilam-iq'(-illi)* N;COMPV(;**ABL-1**).

In this way, neither the current UniMorph design with its non-consideration of order needs to be abandoned nor do we lose information anymore.

This allows to keep existing annotation for nouns intact, as these are currently annotated for their inherent case only. This approach also aligns very well with the impression that multiple case marking is somewhat exceptional, so that 'basic' annotations focus on inherent case.

4. On verbal inflection

In the example in Fig. 1, the Megrelian verb shows head marking for both its syntactic arguments, a first person subject and a second person object. UniMorph allows us to distinguish both clearly, by forming compound features of argument case and argument features, e.g., ARGNO1S for the nominative subject as being first person singular, and ARGDA2S for the dative object as being second person singular (Sylak-Glassman, 2016, p. 13). We discuss problems of argument identification by morphological case for the example of Kartvelian, and suggest an alternative.

4.1. Complex patterns of argument marking

As Fig. 1 shows, verbs in Kartvelian languages can specify grammatical features for *multiple* arguments, so that the agreement information about one (e.g., subject) argument must be clearly distinguished from agreement information about another (e.g., object). This is not specific to Caucasian languages, but does also occur, for example, in Basque, and has been addressed in UniMorph before. The UniMorph solution to the problem is to form compound features, which has the drawback that the UniMorph schema is partially redundant; the following statements are equivalent:

```
has     have     V;3;SG;IND
has     have     V;ARGNO3SG;IND
```

In practice, this problem does not occur, because ARG features are provided for languages where verbs are marked for their arguments. Nevertheless, UniMorph was designed with the intention to project morphological annotations between languages (Yarowsky et al., 2001; Sylak-Glassman et al., 2015a). It is not clear, however, to what extent compound features such as `ARGNO3SG` can be put in *any* relation with the 'regular' features `3;SG` unless additional (language-specific!) assumptions about case morphology and its relation with subjecthood are taken into consideration.

In particular, this is a problem for Kartvelian. It is an established convention for most languages to identify arguments in terms of their grammatical roles ('subject' and 'object'). In Georgian (as well as in Megrelian and other Kartvelian languages), however, the linking between grammatical roles and morphological cases is relatively complex, and the same role can by expressed by different cases, depending on the tense/aspect of the verb, while at the same time, this argument can be marked on the verb by a single element, regardless of the case of the (pro)nominal it refers to.

(8) a. *bavšv-eb-s* *da-∅-malav-s*
child-PL-DAT PV-O3-hide-S3SG

'S/he will hide the children.'

b. *bavšv-eb-i* *da-∅-mala*
child-PL-NOM PV-O3-hide.AOR.S3SG

'S/he hid the children.'

Hence, in the Georgian sentences (8a) and (8b), the 3rd person object is marked by the absence of a prefix (i.e. the 1st and 2nd persons would have been marked). In the future tense, the object constituent receives the dative case, while in the aorist tense it receives the nominative case. Referring to grammatical roles here would be less confusing and more in line with established research.

When populating UniMorph inventories from existing glossed corpora, however, we need to keep in mind that tense may be indicated by *multiple* morphemes which do not have a clear interpretation in a morpheme-based annotation (as for the Megrelian example in Fig. 1). Accordingly, only a language expert can implement a direct mapping between existing morpheme-based annotations (which

refer to grammatical roles) and cases of the corresponding arguments (whose identification depends on identifying the tense feature) – which may create a barrier for creating UniMorph resources.[13]

4.2. A ranking, again

In extension of the ranking-based modeling of multiple case marking, it is possible to generalize over both the case-based and the grammatical-role-based encoding of arguments as well as over compound and regular features for arguments in different languages.

At least since Dowty (1991) and Grosz et al. (1995), the importance of aligned hierarchies of grammatical and semantic roles has been recognized in different communicative functions, and established as such in both linguistics and natural language processing: According to Dowty, a ranking of semantic roles (AGENT > PATIENT > ...) is underlying the assignment of grammatical roles; according to Grosz et al., a ranking of grammatical (or semantic) roles is taken to reflect and to establish discourse salience (Subject > Object > ...) which is closely tied with pronominalization. By extension of this approach, highly salient discourse referents can be expressed by $\emptyset$ pronouns or by verbal inflection, alone, thereby establishing a grammaticalization path from pronouns to verbal inflection (Ariel, 2000). In summary, a ranking of grammatical (semantic) roles is almost universally upheld, and a close relation with verbal morphology is assumed, at least.

Following the UniMorph approach to render grammatical roles with morphological cases, it would thus seem possible to provide a language-specific ranking of morphological cases that represent these cases – or the underlying grammatical roles. Such a ranking, now, can be expressed by numerical indices, as well, with the top-ranked element being assigned the empty index. With a ranking of nominative (subject) over dative (object) over other cases, we can thus develop an alternative representation of Megrelian *kešerxvaduk*, i.e.,

V;...;1;SG;2-1;SG-1
instead of V;...;ARGNO1S;ARGDA2S

This approach elegantly overcomes the asymmetry between compound and individual features, it establishes a principled approach to deal with the assignment of multiple instances of the same feature to (different arguments/heads of) a single word in both the nominal and the verbal domain, and it can be formulated without a priori restrictions to certain grammatical features. UniMorph will thus gain in scalability. Moreover, it eliminates alternative encoding strategies for the same phenomenon, and it even facilitates comparability across languages , as the Megrelian features now overlap with, say, those of its Russian translation *vstrechu*: V;...;1;SG (assuming nominative > other cases). Again, this extension has little impact on most existing UniMorph resources, as ARG features are used for Basque only, so far. For other languages, one would formally need to define a ranking, but default rankings can be posited, too. For accusative languages, the corresponding default ranking would be

nominative > accusative > other

For languages with ergative alignment, the default ranking would be

ergative > absolutive/nominative > dative > other

This hierarchy is suggested to reflect the fact that the ergative case is almost exclusively used for subjects, constituents marked with nominative/absolutive can by subjects or direct objects, and the dative case can be used to mark subjects and both direct and indirect objects. Thus, the hierarchies for both accussative and ergative languages correspond to the hierarchy for grammatical roles:

subject > object > other

5. On the UniMorph format

So far, we discussed possible extensions of the current UniMorph schema that arise from our work on Caucasian and other low-resource languages. However, also the UniMorph file format may represent a hurdle for its application beyond NLP. Although its minimalistic design establishes a high level of interoperability, it seriously limits the usability of UniMorph data sets for linguistic research – and their acceptability for linguists. Therefore, we suggest a workflow to derive the current TSV format from more expressive formalisms that are closer to current practices in language contact studies, language documentation and linguistic research in general.

5.1. Beyond Tab-Separated Values

In a field of research where Interlinear Glossed Text and an elaborate toolchain for its generation and processing (including Toolbox[14] and FLEx[15]) is the state of the art, converting carefully constructed, high-quality morpheme inventories to an incomplete and less interpretable representation poses a problem regarding acceptability and dissemination. The UniMorph format must not be understood as a full-fledged representation formalism, but rather as an interchange format between rich and high-quality language resources on the one hand and morphological generators on the other hand, as developed, e.g., in the context of the SIGMORPHON shared tasks. Moreover, the UniMorph repositories are very likely to get out of sync with the underlying resource, as they are maintained in Github repositories structured according to the same conventions: This means that the source data and its UniMorph extract are maintained at different locations. This maintenance aspect is not that complicated for high-resource languages, as their morphological description is unlikely to evolve greatly in the immediate future. In the context of low-resource languages, however, efforts in language documentation frequently lead

[13] We would like to point out that the current Georgian UniMorph data does *not* include compound features for multiple argument marking, probably for precisely this reason. Furthermore, the Basque UniMorph data (that *does* employ compound features) deviates from the schema by partly using grammatical role labeling (instead of case labeling).

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

[15] https://software.sil.org/toolbox/

to novel insights into the inventory and the function of inflectional morphemes in a language.

As an alternative to the current UniMorph publication model, we propose a formalism and a workflow that allows to embed UniMorph linkings into existing resources, in particular, if these are provided in XML, CSV, JSON, or RDF. The current TSV format can then be retrieved from various types of source data, and the UniMorph repositories can be populated with morpheme inventories in their native representation, avoiding information loss and forks between different versions of the same resource.

```
:sl_3052 a ontolex:LexicalEntry;                  # 1: explicit data structures
    ontolex:canonicalForm
    [ ontolex:writtenRep "xvad" ];                # 2: lemma
    ontolex:otherForm :sl_3052_15.                # 3: link with form(s)
:sl_3052_15 a ontolex:Form;                       # 4: explicit data structures
    ontolex:writtenRep "kešerxvaduk"@xmf;         # 5: word form
    unimorph:hasFeature unimorph:V, unimorph:SG.  # 6: ontology linking, samples
```

Figure 2: Megrelian lemon/RDF sample in Turtle

5.2. Linking UniMorph

Our solution builds on modelling language resources, resp. the linking between them, on the basis of **Linked Data** formalisms. The Linked Data paradigm (Berners-Lee, 2006) postulates rules for the publication and representation of Web resourceswhich facilitate information integration, and thus, interoperability. Data should be represented by means of W3C standards, such as RDF (Resource Description Framework). RDF provides a generic data model based on labeled directed graphs, which can be serialized in different formats. Information is represented by *triples* which consist of a *predicate* (a relation, i.e., a labeled edge) that connects a *subject* (a so-called "RDF resource", i.e., a labeled node) with its *object* (another RDF resource, or a literal, e.g., a string). The RDF resources are represented by URIs, making them unambiguous in the web of data, allowing resources hosted at different locations to refer to each other and thus creating a network of data collections with densely interwoven elements (Chiarcos et al., 2013).

Linked Data has been successfully applied to convert and link language resources (Chiarcos et al., 2012), leading to the emergence of the Linguistic Linked Open Data (LLOD) cloud (Chiarcos et al., 2013; McCrae et al., 2016),[16] a set of linked open language resources for all fields of linguistics, digital philology, natural language processing, the localization industry and the Semantic Web community, tied together by shared vocabularies, the use of reference knowledge bases and links between each other. As such, the lemon/ontolex vocabulary (Cimiano et al., 2016) developed as a community standard for machine-readable dictionaries in the cloud, and its extension to morpheme inventories is currently being discussed. In the context of UniMorph and language documentation, recent proposals to develop vocabularies for Interlinear Glossed Text (Chiarcos et al., 2017) and TSV-based corpus formats (Chiarcos and Fäth, 2017, CoNLL-RDF) are to be mentioned. On the basis of CoNLL-RDF, we developed a tool for the LLOD conversion of the UniMorph format as part of our LLODifier library[17]. We envision a future infrastructure for UniMorph where different source formats are mediated by RDF representations and associated SPARQL scripts. These scripts can then be used to derive the TSV format as currently in use, or – alternatively – can be digested directly (and losslessly) by downstream applications. As an example, we developed converters from FLEx to FLEx RDF (applied to Megrelian), from CoNLL to CoNLL-RDF (applied to Classical Armenian), from ELAN[18] to ELAN RDF (independently from Caucasus studies applied to Old High German), and from TSV to lemon/RDF (applicable to every existing UniMorph dataset).

A key benefit of representing language resources in RDF is that individual items within a resource are identifiable by means of a URI so statements about them can be added easily, e.g. explicit links with an ontology. We provide an OWL2/DL formalization of the UniMorph schema,[19] designed as an Annotation Model in the OLiA architecture (Chiarcos and Sukhareva, 2015). It is thus possible to provide declarative links between individual items in a morphological inventory and the UniMorph ontology:[20] Fig. 2 shows a fragment from the lemon edition of the Megrelian UniMorph inventory, with the original UniMorph features transposed into explicit links to well-defined entities in the web of data (where data consumers can look up the definition, relation to other features, etc.).[21]

5.3. Beyond RDF

RDF technology does, however, not require the source data to be RDF. RDFa (Adida et al., 2015), for example, permits to add typed links to XML documents, which can then be parsed into other RDF serializations. Alternatively, explicit RDF conversion instructions can be attached to XML documents using GRDDL (Connolly, 2007). Similarly, tabular data (as in the current UniMorph format) does not require an explicit conversion: CSV2RDF (Tandy et al., 2015) is a W3C recommendation that allows the *direct* interpretation of tabular data as RDF – and thus enables its linking with, say, the UniMorph ontology. For other formats designated converters are provided, for example, as part of the LLODifier library.

Such conversions from various source formats merely require (a) an indication of their original format (TSV, RDF, XML – for W3C-supported formats), resp., the converter (for other formats), and (b) one SPARQL Update script per source format to guarantee conformancy with common specifications, e.g., to resolve feature abbreviations into links against the UniMorph ontology. The latter, however,

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

[17] https://github.com/acoli-repo/LLODifier/tree/master/unimorph.

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

[19] http://purl.org/olia/owl/experimental/unimorph

[20] Such links can be auto-generated from abbreviations, cf. https://github.com/acoli-repo/LLODifier/blob/master/unimorph/link-and-load-FEATS.sparql.

[21] This linking mechanism can also be used to map from an existing annotation scheme into UniMorph, as currently implemented, e.g., from the PROIEL schema to Classical Armenian.

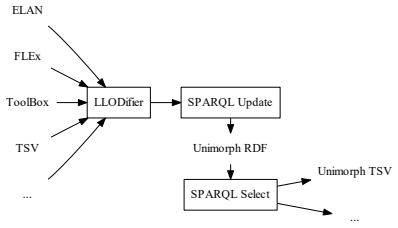

Figure 3: Suggested pipeline for converting data to UniMorph RDF

are optional, and invariant for each type of source format.

5.4. Back to TSV

With RDF data or an RDF interpretation of source data in place, a TSV file can then be automatically generated using a SPARQL SELECT statement, e.g., from a lemon RDF file:

```
SELECT ?word ?lemma ?feats
WHERE {
   ?form a ontolex:Form; writtenRep ?word.
   ?lexEnt ontolex:otherForm ?form;
           ontolex:canonicalForm/ontolex:writtenRep
           ?lemma.
    {
    SELECT ?word (GROUP_CONCAT(?feat; separator=";")
    AS ?feats)
    WHERE {
        ?word unimorph:hasFeature/unimorph:hasLabel
        ?feat
    } GROUP BY ?word
    }
}
```

The output format of this query is a table and its TSV serialization can be directly fed into existing UniMorph-based tools.

The use of RDF for data conversion and SPARQL for data transformation and querying thus facilitates the development of a technical infrastructure for UniMorph which allows the community to grow beyond the limitations imposed by the crippling TSV format it is currently based on, an achievement which would be most welcome to linguists, researchers and NLP engineers working on low-resource languages.

6. Summary and outlook

We describe the creation of UniMorph resources for languages in the Caucasus region, including Megrelian, Khinalug, and Classical Armenian, which are published under an open license via our UniMorph fork[22] and which are to be integrated with the main UniMorph infrastructure (in case our suggested modifications meet community approval), thereby increasing the coverage of languages from the Caucasus area in UniMorph.

We discussed a number of peculiarities of these languages and potential conceptual difficulties in the application of the UniMorph scheme to them and other languages. As a result, we suggest the following extensions to the UniMorph schema:

- human-readable labels for LGSPEC features, e.g., LGSPEC-TM instead of LGSPEC4 for Megrelian,
- a ranking-based numerical scheme to represent iterative features in nominal inflection,
- a ranking-based numerical scheme to encode multiple arguments of polyvalent verbs in head-marking languages, and
- the postulation of a default ranking for verbal arguments, as well as the possibility to posit language-specific rankings.

In addition, we discuss the UniMorph TSV format and criticize its limited expressiveness which creates a gap between its uses in NLP and potential users of UniMorph technology or providers of UniMorph data in linguistics. We thus suggest to complement the existing UniMorph workflow with support for additional source formats on grounds of Linked Open Data technology. For this purpose, we provide converters for UniMorph TSV, FLEx, ELAN and other formats to RDF,[23] a SPARQL query for the generation of UniMorph TSV out of RDF and an RDF/OWL edition of the UniMorph schema that we provide as part of the Ontologies of Linguistic Annotation.[24]

The combination of these resources allows us to derive Unimorph TSV files from various source formats, and our UniMorph fork provides not only TSV files, but also Makefiles and associated resources. For the future, however, one may consider to follow a streamlined approach and develop a uniform UniMorph representation in RDF, which can be derived from resource-specific RDF representations and mediate between these and the current UniMorph TSV representation as illustrated in Fig. 3.

One key advantage of a future RDF vocabulary of UniMorph data in comparison to TSV data would be that additional data can be added as needed, without affecting its processability. In particular, it may preserve *any* information from the original, resource-specific RDF – just cleanly separated in a distinct namespace. Such a UniMorph vocabulary could build, for example, on existing community standards such as lemon (Cimiano et al., 2016), as illustrated in Fig. 2.

7. Acknowledgements

This research was conducted in the context of the research group 'Linked Open Dictionaries' (LiODi, 2015-2020, Goethe University Frankfurt, Germany). LiODi is funded by the Federal Ministry of Education and Research and its activities are centered around developing innovative methodologies for utilizing dictionaries and corpora for language contact studies, with a particular focus on applications of Linguistic Linked Open Data. LiODi consists of researchers from applied computational linguistics and empirical linguistics, with a focus on language contact in the Caucasus area.

8. Bibliographical references

Adida, B., Birbeck, M., McCarron, S., and Herman, I. (2015). RDFa Core 1.1 - Third Edition. Syntax and

[22] https://github.com/acoli-repo/unimorph

[23] https://github.com/acoli-repo/LLODifier

[24] http://purl.org/olia/owl/experimental/unimorph

processing rules for embedding RDF through attributes. Technical report, W3C Recommendation.

Ariel, M. (2000). The Development of Person Agreement Markers: from Pronouns to Higher Accessibility Markers. In Michael Barlow et al., editors, *Usage-based models of language*, pages 197–260. CSIL Publications, Stanford, CA.

Berners-Lee, T. (2006). Design issues: Linked data. Technical report.

Boeder, W. (1995). Suffixaufnahme in Kartvelian. In Frans Plank, editor, *Double Case: Agreement by Suffixaufnahme*, pages 151–215. Oxford University Press, Oxford.

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*, pages 74–88, Cham. Springer.

Chiarcos, C. and Sukhareva, M. (2015). OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Christian Chiarcos, et al., editors. (2012). *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*. Springer, Berlin, Heidelberg.

Chiarcos, C., Moran, S., Mendes, P. N., Nordhoff, S., and Littauer, R. (2013). Building a Linked Open Data cloud of linguistic resources: Motivations and developments. In *The People's Web Meets NLP. Theory and Applications of Natural Language Processing*, pages 315–348. Springer, Berlin, Heidelberg.

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C., Wichers Schreur, J., and Nevskaya, I. (2017). LLODifying Linguistic Glosses. In *International Conference on Language, Data and Knowledge*, pages 89–103, Cham. Springer.

Cimiano, P., McCrae, J., and Buitelaar, P. (2016). Lexicon Model for Ontologies. Technical report, W3C Community Report.

Connolly, D. (2007). Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Technical report, W3C Recommendation.

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational linguistics*, 21(2):203–225.

Haspelmath, M. (2007). Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1):119–132.

Hübschmann, H. (1906). Armeniaca. *Indogermanische Forschungen*, 19(1):457–480.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris. European Language Resources Association (ELRA).

Klimov, G. A. (1994). *Einführung in die kaukasische Sprachwissenschaft*. Buske Helmut Verlag GmbH.

Kracht, M. (2003). Against the feature bundle theory of case. In Ellen Brandner et al., editors, *New Perspectives on Case Theory*, pages 165–190. CSLI Publications, Stanford, CA.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., et al. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2435–2441, Paris. European Language Resources Association (ELRA).

Nivre, J. et al. (2015). *Universal Dependencies 1.2*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Plank, F. (2014). *Double case: Agreement by Suffixaufnahme*. Oxford University Press.

Radkevich, N. (2010). On Location: The Structure of Case and Adpositions. Ph.D. thesis, University of Connecticut, Storrs, CT.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015a). A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging. In Cerstin Mahlow et al., editors, *Systems and Frameworks for Computational Morphology*, pages 72–93, Cham. Springer.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015b). A Language-Independent Feature Schema for Inflectional Morphology. In Chengqing Zong et al., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing. Association for Computational Linguistics.

Sylak-Glassman, J. (2016). The Composition and Use of the Universal Morphological Feature Schema (Unimorph Schema). Technical report, Department of Computer Science, Johns Hopkins University. working draft, v.2.

Tandy, J., Herman, I., and Kellogg, G. (2015). Generating RDF from Tabular Data on the Web. Technical report, W3C Recommendation.

Wegner, I. (1995). Suffixaufnahme in Hurrian: Normal Cases and Special Cases. In Frans Plank, editor, *Double Case: Agreement by Suffixaufnahme*, page 136–147. Oxford University Press, Oxford.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology (HTL)*, pages 1–8, Stroudsburg, PA. Association for Computational Linguistics.

9. Language Resource References

Haug, D. and Jøhndal, M. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations.

In Caroline Sporleder et al., editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Marrakech.