Universal Morphology zwischen Sprachtechnologie und Sprachwissenschaft: Sprachressourcen für Kaukasussprachen

Chiarcos, Christian
chiarcos@informatik.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Donandt, Kathrin
donandt@informatik.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Ionov, Maxim
ionov@informatik.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Rind-Pawlowski, Monika
pawlowski@lingua.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Sargsian, Hasmik
Sargsyan@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Wichers Schreur, Jesse
wichersschreur@em.uni-frankfurt.de
Goethe-Universität Frankfurt, Deutschland

Hintergrund

Das Projekt `Linked Open Dictionaries' (LiODi, 2015-2020) ist eine vom Bundesministerium für Bildung und Forschung (BMBF) finanzierte eHumanities-Forschungsgruppe, die daran arbeitet, einen nutzerorientierten Zugang zu LOD-Technologien in den Sprachwissenschaften zu entwickeln und diesen in Einzelstudien zum Sprachkontakt im Kaukasus zu demonstrieren. Ein wichtiges Element dafür sind Lehnwortuntersuchungen, und in morphologisch reichen Sprachen ist es möglich, dass eine flektierte Form Gegenstand der Entlehnung war. Diese automatisch gestützt generieren und identifizieren zu können, erfordert einen morphologischen Generator, der auch über unvollständige Daten hinweg generalisieren, und beispielsweise Paradigmen komplettieren kann. Hierfür stellt die Universal Morphology derzeit Standardressourcen bereit, auf die hin Softwareimplementierungen optimiert werden, etwa im Rahmen aktueller SIGMORPHON Shared Tasks (Cotterell et al., 2016).

Universal Morphology (*Unimorph*, http://unimorph.github.io/) ist ein aktuelles Communityproject zur Erfassung und automatischen Generierung der Flexionsmorphologie unterschiedlichster Sprachen. Ziel ist sowohl die Entwicklung von Vollformenwörterbüchern, deren Einsatz zur Annotation, weshalb Unimorph auf Kompatibilität mit den Universal Dependencies (http://universaldependencies.org/) angelegt ist, aber auch der Aufbau einer Referenzressource zur Entwicklung morphologischer Analyse- und Generierungskomponenten innerhalb der Sprach *technologie*. Die sprach *wissenschaftliche* Nutzung jedoch steht bislang aus und bestimmt daher den Fokus unseres Beitrages. Unimorph-Ressourcen und -Technologien sind dabei eine höchst willkommene Ergänzung unserer Arbeit, ihre praktische Anwendung des Schemas auf das Kaukasusgebiet erweist sich jedoch als problembehaftet.

Der **Kaukasus** ist für die Diversität seiner Sprachen bekannt, die oftmals eurozentristische Ansichten traditioneller Sprachwissenschaften infrage gestellt haben, und sich daher sehr gut zur Prüfung von linguistischen Modellen mit universellem Anspruch eignen. Viele Kaukasussprachen sind bedroht, die meisten (mit Ausnahme von Georgisch, Armenisch und Albanisch/ Udi) wurden erst in jüngerer Vergangenheit verschriftlicht. Allen gemeinsam ist ein großer Lehnwortschatz (u.a. aus dem Iranischen, Türkischen, Russischen) und Lehnbeziehungen untereinander. Alle Kaukasussprachen sind sprachtechnologisch schlecht erschlossen, hier betrachten wir daher aktuelle Ansätze zur Schaffung von sprachübergreifenden (`universellen‘) morphologischen Annotationen im Rahmen des *Unimorph*-Projektes. Wir berichten Ergebnisse zu unserer Arbeit zu Batsbi (nakh-dagestanisch), Mingrelisch (kartvelisch), Khinalug (nakh-dagestanisch) und Armenisch (indoeuropäisch). Auf dieser Basis diskutieren wir behutsame Erweiterungen von Unimorph, um dessen Anwendbarkeit für die kaukasischen Sprachen im Besonderen und für Sprachdokumentationsdaten im Allgemeinen zu ermöglichen.

Unimorph für Sprachdokumentation im Kaukasus?

Schema und Schemaerweiterungen

Unimorph verwendet ein **TSV-Format**, d.h., eine Liste von tab-separierten Einträgen für jeweils Wortform, Lemma und Unimorph-Tags. Letztere sind *nicht qualifizierte* Merkmale, durch Semikolon getrennt und *unsortiert*. Der Eintrag für deutsch „(ich) treffe (dich)“ wäre beispielsweise

```
treffen  treffe  V;IND;PRS;1;SG
```

Der Eintrag für mingrelisch *kešerxvaduk* (`Ich werde dich treffen') besitzt folgende Glosse:

```
kešerxvaduk
ke-   še-   r-    xvad  -u    -k
AFF   PV    O2SG  meet  TM    S1SG
```

In Unimorph wird diese Analyse wie folgt repräsentiert:

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

Das mingrelische Verb kongruiert mit beiden syntaktischen Argumenten: dem Subjekt (1S, Nominativ) und dem Objekt (2S, Dativ), für die **zusammengesetzte Merkmale** gebildet werden, die ein Argument mit dessen Person, Numerus usw. zusammenstellt; z.B. werden hier ArgNo1S für „Nominativargument=1. Person Singular“ und ArgDa2S für „Dativargument=2. Person Singular“ aufgeführt. Ein methodisches Problem ist, dass das Unimorph-Schema dieselbe Information hierbei in unterschiedlicher Weise ausdrückt, wie im Vergleich deutlich wird: mingrelisch ArgNo1S entspricht deutsch 1;SG. Da der Zusammenhang zwischen Kasus und grammatischen Rollen für das Deutsche nicht explizit definiert ist, gibt es keine Möglichkeit, diese automatisch als äquivalent zu interpretieren. Leider erlaubt es Unimorph zudem nicht, herkömmliche Terminologie zu verwenden, in der beide Argumente bzgl. ihrer syntaktischen Rollen (`Subjekt' und `Objekt') beschrieben werden, sondern zieht stattdessen die Kasusmorphologie heran. Dies ist insofern problematisch, als der Kasus im Verb nicht morphologisch realisiert ist, und Argumente im Satz nicht (pro)nominal realisiert werden müssen. Konventionell werden statt dessen grammatische Rollen verwendet.

Ähnlich zu (mehreren Argumenten von) Verben gibt es Nomina mit **mehrfacher Kodierung derselben Merkmalskategorie** (v.a. Kasus), die mit Unimorph nicht behandelt werden können. In der sog. *Suffixaufnahme* spezifizieren adnominale Elemente *neben* ihrem inhärenten Kasus auch Agreement-Merkmale ihres Kopfnomens, z.B. durch Wiederholung von dessen Kasusmorphologie. Dies wurde ursprünglich für Georgisch beschrieben, gilt aber als verbreitet im Kaukasus. Bedauerlicherweise kann diese Information in Unimorph nicht positional kodiert werden, sondern erfordert eine Erweiterung des Label-Inventars. Deshalb schlagen wir die Einführng numerischer Indizes in der nominalen Morphologie vor, wobei der inhärente Kasus nicht bezeichnet wird, der Kasus des direkten Kopfes durch Anhängen von -1 an das Feature-Label, der Kasus von dessen Kopf durch -2, usw.

Diese **numerischen Indizes** sind auch auf die verbale Domäne übertragbar. Gegeben etablierte Hierarchien grammatischer Rollen bzw. der zugeordneten Kasus, kann die bisherige Verbundmarkierung multipler Argumente durch eine Indizierungsstrategie ersetzt werden, die sich auf diese bezieht, und bei der das höchstrangige Element (z.B. das Subjekt) unbezeichnet bleibt, während andere Argumente nach ihrer Stellung im Ranking gekennzeichnet werden. Eine alternative Repräsentation des mingrelischen *kešerxvaduk* wäre also

```
V;... 1;SG; 2-1;SG-1
```

Dies korrigiert auch die Asymmetrie zwischen zusammengesetzten und individuellen Merkmalen. Auch die Zuschreibung mehrerer Merkmale einer Kategorie kann nominal und verbal einheitlich gehandhabt werden, und die Vergleichbarkeit über Sprachen hinweg wird vereinfacht.

Datenformat und Alternativen

Ein zweites Problem ist das in Unimorph verwendete, nicht erweiterbare TSV- **Format**, das gegenüber den in der Sprachdokumentation üblichen Softwarelösungen (FLEx, Toolbox, ELAN) aber nur *stark eingeschränkte* Informationen bereitstellt: Im Vergleich zur wörterbuchgestützten Interlinearglossierung stellen Morphem-Inventorien in unvollständigen und weniger gut interpretierbaren Repräsentationen ein Akzeptanzproblem dar. Daher sollte Unimorph nicht als eigenständiger Formalismus verstanden werden, sondern als Austauschformat zwischen reichen und hochwertigen Sprachressourcen auf der einen Seite und morphologischen Generatoren auf der anderen.

Allerdings sind *Format und Speicherort* festgeschrieben, so dass zugrundeliegende Ressourcen an anderen Orten gespeichert und gepflegt werden müssen, und Ergänzungen aus der Sprachdokumentationsarbeit womöglich nicht eingepflegt werden. Wir schlagen daher vor, das jetzige Format nur bei Bedarf zu generieren. Der Schlüssel hierzu liegt darin, die Quellformate gemäß W3C-Standards zur Ressourcentransformation auf einheitliche RDF-Datenstrukturen nach lemon (https://www.w3.org/2016/05/ontolex/) zu mappen und mit Hilfe der Anfragesprache SPARQL das derzeitige Tabellen-Format zu erzeugen. Das Repository enthält dann für jede Sprache die (a) *vollständigen* Daten, und (b) ein standardisiertes Mapping auf lemon. Die TSV-Generierung ist nicht ressourcenspezifisch. Der Gebrauch von RDF-Technologien für Datenkonversion und Abfrage kann so die Entwicklung einer technischen Infrastruktur

für Unimorph ermöglichen, die es erlaubt, über die Grenzen des TSV-Formats hinauszuwachsen, wovon SprachwissenschaftlerInnen, ForscherInnen und NLP-IngenieurInnen, die mit *low-resource*-Sprachen arbeiten, profitieren könnten.

Die Integration mit gängigen Annotationswerkzeugen kann hierbei auf von uns entwickelten RDF-Konvertern für FLEx, Toolbox und weitere Formate aufsetzen (Chiarcos et al., 2017, https://github.com/acoli-repo/LLODifier).

Zusammenfassend plädieren wir für die Einführung numerischer Indizes für verschiedene Argumente polyvalenter Verben und rekursive Merkmale in der Nominalflexion in Unimorph. Für die bessere Integration von existierenden Ressource aus der Sprachdokumentation insgesamt schlagen wir zudem eine Erweiterung der unterstützten Formate und einen einheitlichen Zugriff auf diese auf Basis von RDF-Technologien vor, so dass das jetzige Unimorph-Format nicht mehr von den zugrundeliegenden, reicheren Quelldaten separiert wird, sondern bedarfsabhängig daraus generiert wird. Sind beide Mängel behoben, steht einer sprachwissenschaftlichen Nutzung von Unimorph hinsichtlich der Sprachkontaktforschung im Kaukasus nichts mehr entgegen.

Bibliographie

Chiarcos, Christian / Ionov, Maxim / Rind-Pawlowski, Monika / Fäth, Christian / Wichers Schreur, Jesse / Nevskaya, Irina (2017): “LLODifying linguistic glosses” in: *Proceedings of the First International Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 2017*. Springer (Lecture Notes in Artificial Intelligence (LNAI)), 89-103 https://doi.org/10.1007/978-3-319-59888-8_7 [letzter Zugriff 14. Januar 2018]

Cotterell, Ryan / Kirov, Christo / Sylak-Glassman, John / Yarowsky, David / Eisner, Jason / Hulden, Mans (2016). “The SIGMORPHON 2016 Shared Task Morphological Reinflection” in: *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, Berlin, Germany, August, 2016*. Association for Computational Linguistics, 10-22 http://anthology.aclweb.org/W16-2002.pdf [letzter Zugriff 14. Januar 2018]

Sylak-Glassman / John (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, https://unimorph.github.io/doc/unimorph-schema.pdf [letzter Zugriff 14. Januar 2018]