Meeting 24/07/2020

Attendees

1. Eleanor Chodroff
2. Reut Tsarfaty
3. David Yarowsky
4. Ekaterina Vylomova
5. Arturo Oncevay
6. Ali Salehi
7. Mans Hulden
8. Omer Goldman
9. Ryan Cotterell
10. Antonis Anastasopoulos
11. Tiago Pimentel

1) **Data storage: what would be the best place to store/share data while we are working on it?**

   Option 1: Google Drive
   Option 2: Git (might not be suitable for all of us)

   We've decided to try Google Drive with sync with a single private git (as a backup)

   We agreed to have a single shared folder with subfolders corresponding to working groups.

   The root folder will also contain a shared file with Q&A/discussions for the whole group:
   https://drive.google.com/drive/folders/1FuvWuRm4Ypr96uXrRJtFlejvKZ0Txc-c

   Subfolders will contain all language data files + gdocs that document feature conversion decisions, etc (maybe create subfolders such for “raw” and “ready” data?).

   Kat: I provided refs to some of them in the UniMorph spreadsheet:
https://docs.google.com/spreadsheets/d/1OA3m_kTnhYMZK762x1SiWy7wMSijxtTfAyDXB9V3wGY/edit?usp=sharing

Kat: We initially put some language data here:
https://drive.google.com/drive/folders/1Y-r7NQJ1jtUgsfYxrnT4oL0R3DZn6kym
Shall we move them?

2) UniMorph release process: follow UD (X.Y)?

Two types of extension: data + schema

Two times a year:
.Y -- no schema changes (just data)
X. -- schema changes

3) Documentation: Web interface

Q.: how to communicate the schema? Now we just have a "frozen" version pdf which is hard to use.
pdf--> lightweight documentation easy to navigate

Important: use Leipzig style + provide some extra description for a phenomenon

There're some inconsistencies to be addressed: e.g. grammatical cases across languages

We've agreed to have a student to focus on it
Kat: I am happy to help with html/web interface part

Meeting 22/07/2020

Attendees

1. Eleanor Chodroff
2. Clara Vania
3. Maria Ryskina
4. Elizabeth Salesky
5. Salam Khalifa
6. Kate Lindsey
7. Antonis Anastasopoulos
8. Carmel O'Shannessy
9. Ryan Cotterell
10. Marc Canby
11. Ekaterina Vylomova

1) Annotation efforts (and UniMorph)

1.1 UniMorph schema/annotation and working groups

1. UniMorph schema: https://unimorph.github.io/doc/unimorph-schema.pdf (all features (tags) are provided there, it's not an exhaustive list, extra features can be added upon agreement within UniMorph community)

   This year we also added a script: https://github.com/unimorph/um-canonicalize
   It checks that all features are in some canonical order and no two values (say,PST;FUT) are assigned to the same feature (e.g., Tense). *This introduced some complexity in the annotation of case compounding (that is present in some Australian + Uralic languages).*
   All features and values are listed in https://github.com/unimorph/um-canonicalize/blob/master/um_canonicalize/tags.yaml

2. Current language data can be accessed here: https://unimorph.github.io/ (some language-specific features are annotated as LGSPEC[0-9])

3. This year data (not in UniMorph yet):
   This year's new shared task data is not yet in UniMorph and can be (partially) accessed here: https://github.com/sigmorphon2020/task0-data/ (at the moment there's an upper threshold for the number of <lemma; tags; form> samples per language (100k); we will share full data shortly)
   (we have a slightly updated version of the tags.yaml file with a few extra (not yet added to UM) features here: https://github.com/sigmorphon2020/task0-data/blob/master/tags.yaml)

   We have also added a few other scripts for data consistency checks that I will share during the weekend.

4. A list of all languages we are working on (some working groups are provided there; some are missing -- feel free to create them): https://docs.google.com/spreadsheets/d/1OA3m_kTnhYMZK762x1SiWy7wMSijxtTfAyDXB9V3wGY/edit?usp=sharing
   Each language is assigned to a working group (that is organized based on language families/typological similarities). It might not be the best (e.g. we can split Papua New Guinea and Australian languages). Suggestions are welcome!

   Each working group will have a Google group to discuss annotation: grammatical features, transcription, etc.
   -- Check our annotation for languages that are already in UniMorph; suggest any modification; discuss it within a group
   -- Discuss any extra features/values that should be added to the schema. If the group agrees on some feature/value, they should share the suggestion with the unimorph community, and make sure we agree that we should add it to the schema

-- Agree on transcription format, etc
-- (Optionally, when possible) Annotate cognates, IPA, provide English translations

1.2 For new languages (conversion):

All ***new*** languages are listed here
https://docs.google.com/spreadsheets/d/1NtG7jOHGqj4JrJJHjVYPZQ1OIr25_TgS5LDSjKsvSTs/edit?usp=sharing
(together with their sources such as finite-state analyzers, grammar books, inflection tables).

We should first extract the data in a language-specific format and then start converting it into the UniMorph schema. Working groups will focus on the consistency of the conversion (so far, we didn't do that, unfortunately). Would be great to have a shared document where all decisions are listed.

2) Shared Task 2021

As Maria suggested, we could focus on different aspects of generalization and have 2 subtasks/tracks: 1) generalization across language families (as we did in 2020); 2) (cognitively plausible) generalization wrt human ("wug-test")

The data for (1) will come from the current annotation iteration.

For (2): We will only focus on some languages with clean data. We will additionally ask native speakers(linguists?) to participate in "wug-test" in their languages.
Kat (is it kinda contextual inflection?): Sample from UM --> transform tags in UD --> Find corresponding tag combination in UD → Extract sentential context --> replace a target word with a generated "wug" --> ask annotators to propose forms (how about inherent features? We should provide them then or somehow remove ambiguity)

Francis (from emails): consider https://github.com/ftyers/vardial-shared-task (the data/task)?

Meeting 20/07/2020

Attendees

1. Eleanor Chodroff
2. David Yarowsky
3. Nizar Habash
4. Adam Wiemerslage

5. Clara Vania
6. Claudia Borg
7. Ryan Cotterell
8. Ekaterina Vylomova
9. Ling Liu
10. Mans Hulden
11. Salam Khalifa
12. Winston Wu
13. Xingyuan Zhao
14. Zoey Liu
15. Miikka Silfverberg

How to modularize the UniMorph annotation effort?

The schema (2016):
http://www.unimorph.org/doc/Sylak-Glassman_2016_-_UniMorph_Schema_User_Guide.pdf

In terms of annotation/features: most importantly, we should try to follow Leipzig convention

Paradigm completeness: in UM there's no requirement for complete paradigms (can be partial paradigms)

Would be great to have more consensus (on feature conversion/orthography/transcription/etc) at least on family level
(For features, we previously had lang-specific conversion like https://docs.google.com/spreadsheets/d/1xKg0JRgHqWr6ohbl0bhNKZU2lBfsIp9sPMWeUEjcBeM/edit?usp=sharing; still many problems with case system conversion)

Important: some languages in UniMorph weren't checked by native speakers, ideally it should not be the case (Kat: and there are issues due to incorrectly extracted paradigm tables from Wiktionary in Adyghe, Yiddish, and some other languages)

Several decisions to be made:

-- Add phonology as an extra option?

-- Different spellings (non-standardized orthography); addressing political/cultures issues (which orthography to use)

-- Annotating clitics (follow LORELEI?)

-- Word vs. multi-word sequences (boundaries of morphology)
(also: Kat: in Russian there're phrases that require dependency parsing)

-- Interface with UD (CC Dan Zeman)

Turkish: mark/annotate morpheme -- feature correspondence

Working Groups

Guidelines/collaboration/discussion media: for each language family we should provide an updated specification. It can be:
-- github (might be problematic for some linguists)
-- google docs + https://sites.google.com/a/nyu.edu/coda/general-rules (via “Awesome Tables”)
https://camel.abudhabi.nyu.edu/madar/ (as Nizar Habash suggests)
-- (we are planning to have such interface within) unimorph platform

TODO: Take a few language families and create a template for a specification (agree on terminology/unification/political issues resolved (which orthography to use, etc))

We can start with:

Semitic (Nizar Habash, Salam Khalifa):
https://groups.google.com/forum/#!forum/unimorph-semitic
Turkic (Eleanor Chodroff):
Balto-Slavic (Ekaterina Vylomova Ryan Cotterell):https://groups.google.com/forum/#!forum/unimorph-slavic
Uralic (Mans Hulden): https://groups.google.com/forum/#!forum/unimorph-uralic
//Kat: can we at some point decide about case compounding?
Austronesian (Clara Vania, Ryan Cotterell):
https://groups.google.com/forum/#!forum/unimorph-austronesian
Romance ():https://groups.google.com/forum/#!forum/unimorph-romance

Timeline:

Hard deadline for all data: end of Dec

UniMorph's citation and publication strategy (should be friendly for field linguists+theoretical ling + comput ling)

Language-specific vs. family (Follow the format like http://www.oto-manguean.surrey.ac.uk/Info/Cite ?)

We mostly agreed that it should be a living (Edition 1,2,.., n) handbook (“UniMorph: Universal Morphology”) as a long-term goal

With an intro chapter on handing of clitics, case typology, etc. (from computational and linguistic chapters)

Chapter per family (discussion of scripts, tools, benchmarks)

Publisher (preferably allows editions)---?
How about https://langsci-press.org/ (from Kat)?

+ Mans also suggested a publisher (what was the name?)

We can also publish language family-specific papers in linguistics/comp. linguistics conferences/journals

Or (arxiv + workshop (for LFs) + compiled into a handbook)

Kat: I was thinking we could also publish papers based on error analysis for a target language families, similar to:

https://www.aclweb.org/anthology/K19-1014/

or

http://jlm.ipipan.waw.pl/ojs/index.php/JLM/article/viewFile/244/238

or

https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1140&context=scil

Shared task

-- Focus on clean and diverse data (so it can also be used by linguists)

-- Need to discuss what aspects of language we should reflect in the data (to make it more useful for community)

Meeting 22/10/2020

Attendees

1. Ben Ambridge
2. Ekaterina Vylomova
3. Ryan Cotterell
4. Brian Leonard
5. Tiago Pimentel

What should be the procedure for sampling a good wug set for the inflection task?

Assuming that we have a large set of wug forms to work with, we decided that the best procedure would be to run a UM-trained language model on all the wugs and select those with the highest prediction entropy, as a way of finding forms where the choice of inflection is the least deterministic.

We discussed the issue of false negatives, that is, the possibility of a situation where the model is highly confident about a given inflection, but a human would find the decision much more contentious. These may be due to unpredictable phonological effects, or due to semantic factors like animacy that can't be accounted for in the data. We may include a random distribution of forms in the sample, as a way of trying to find these outliers by casting a wide net.

We decided to run a rating task in addition to the elicitation task. The elicitation task has the benefit of being a more naturalistic context, and is more modular in that it isn't tied to the output of our specific model. The rating task, in which subjects will give continuous likert ratings to the inflected wug forms our model outputs, will allow us to collect info on a large number of wug inflections that the subjects may never have produced independently in an elicitation task.

We decided to begin with four languages: English, German, Portuguese, and Russian.