Visualization of Morphology Datasets

Ziying Zhang[*] Lan Sang[†] Ling Liu[‡]

Abstract

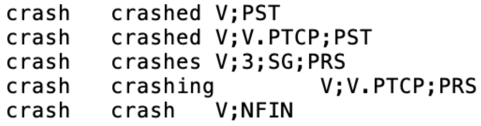
The Unimorph project and the Universal Dependencies project provide two multilingual data resources commonly used for computational morphology. Each project has annotated data of various sizes for around 100 languages of different morphological typology and various language family. More languages are being annotated under each project. Currently, each project lists their languages in tables where they provide basic information about the language and its annotated data. Though these two projects both aim at providing universally annotated data for cross-lingual studies and their data are complementary to each other, neither of them provides information about the language in the other project. If we could get the information about the language in both of the two data resources at one place, it will be a very convenient reference. In addition, morphological complexity is closely related to the amount of training data needed for a good morphological model. An analysis of the morphological complexity as well as whether the language is low-resourced in the dataset or not can provide people with an overview of data availability in the two projects for tasks in computational morphology. Therefore, for our final project, we created a website with visualizations for the data information in the two resources and analyzed the data in Unimorph as to their morphological complexity and low-resource status. This website with visualizations help us to learn about the two resources, and the findings are discussed in the discussion section.

Index Terms: Computational morphology—Visualization—Multiple languages—UnimorphUniversal Dependencies; Computational morphology—Visualization—Morphological complexity; Computational morphology—Visualization—Low-resource Languages;

1 Introduction

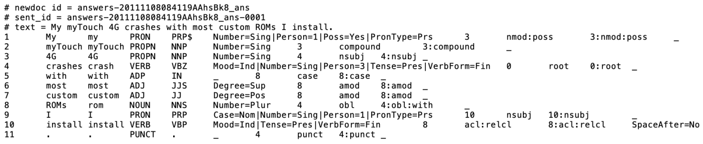
Morphology is the study of word structures. Computational morphology uses computational methods to facilitate and expand morphological studies as well as tackling natural language processing (NLP) problems involving word structures. The Unimorph project (UM) and the Universal Dependencies project (UD) are two complementary resources where people can get annotated data in multiple languages for tasks in computational morphology.

The Unimorph project [16] provides lexicon data, i.e. words are annotated in the format of an inflection table where there is no sentential context. This project now has annotated data for 110 languages and has 52 languages in the process of being annotated according to Unimorph schema. Figure 1 is an example of annotated data in UM. Each group of words like this is called a paradigm, or inflection table. The first column is called dictionary form of the word, the second column is inflected forms of a word, and the third column is morphosyntactic descriptions (MSDs) which describes the morphosyntactic or functional meaning the inflected form adds to the dictionary form in the same row. This is a resource for computational morphology tasks that don't involve sentential context.



```
crash   crashed V;PST
crash   crashed V;V.PTCP;PST
crash   crashes V;3;SG;PRS
crash   crashing        V;V.PTCP;PRS
crash   crash   V;NFIN
```

Figure 1: UM example: Paradigm for English verb "crash"



```
# newdoc id = answers-20111108084119AAhsBk8_ans
# sent_id = answers-20111108084119AAhsBk8_ans-0001
# text = My myTouch 4G crashes with most custom ROMs I install.
1       My      my      PRON    PRP$    Number=Sing|Person=1|Poss=Yes|PronType=Prs      3       nmod:poss       3:nmod:poss     _
2       myTouch myTouch PROPN   NNP     Number=Sing     3       compound        3:compound      _
3       4G      4G      PROPN   NNP     Number=Sing     4       nsubj   4:nsubj _
4       crashes crash   VERB    VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root    0:root  _
5       with    with    ADP     IN      _       8       case    8:case  _
6       most    most    ADJ     JJS     Degree=Sup      8       amod    8:amod  _
7       custom  custom  ADJ     JJ      Degree=Pos      8       amod    8:amod  _
8       ROMs    rom     NOUN    NNS     Number=Plur     4       obl     4:obl:with      _
9       I       I       PRON    PRP     Case=Nom|Number=Sing|Person=1|PronType=Prs      10      nsubj   10:nsubj        _
10      install install VERB    VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        8       acl:relcl       8:acl:relcl     SpaceAfter=No
11      .       .       PUNCT   .       _       4       punct   4:punct _
```

Figure 2: UD example: Annotation for English sentence "My myTouch 4G crashes with most custom ROMs I install."

The Universal Dependencies project [12] provides corpus data, i.e. articles are selected for annotation which are divided into sentences where each word is annotated. The latest version of this project data is version 2.5, which now has annotated data for 90 languages, and there are 16 languages in the process of being annotated according to Universal Dependencies schema. Figure 2 is an annotated sentence example from UD, where the second column provides words as they are in the sentence, the third column is the dictionary forms of the corresponding word whose morphosyntactic meaning are annotated in column four of part-of-speech information and columns five and six of more detailed MSDs. This is a resource for computational morphology tasks that require a sentential context.

Currently, both of these two projects list their languages in alphabetical order of the language name in tables, where basic information about the language and its data is listed. The two projects are similar in that they both aim at providing data annotated with the universal schema to promote cross-lingual studies and NLP tasks. They are complementary in that one is lexicon annotation and the other is corpus annotation. However, there is no place where users can get information about data in these two resources together. Users have to go to the website for each project to look for whether the language has data available in each and find data information in the tables separately. This motivated us to conduct this final project to build a website with visualizations about data information in UM and UD to provide users a convenient reference.

Data are critical for developing good systems. Related to this in computational morphology are two concerns: (1) How complex the morphological system of a language is (i.e. morphological complexity). In general, the more complex the morphological system is, the more data we need. (2) Whether we have abundant annotated data (i.e. low-resource status). Languages with little data available are called low-resourced languages. Information about these two concerns is usually very helpful for people doing computational morphology. Data information in the UM and UD, especially UM, can indicate the morphological complexity of languages in different aspects, and the data size can indicate the low-resource status of the language as to the coverage of the two projects. This motivated us to visualize and analyze the data in UM and UD, especially UM, from these two aspects.

[*] e-mail: zizh8550@colorado.edu
[†] e-mail: lan.sang@colorado.edu
[‡] e-mail: ling.liu@colorado.edu

2 Related work

UM and UD are two common resources for computational morphology. There are a lot of tasks and publications on computational morphology using data from these two resources, including [3, 7–9, 15] etc. There are tools to convert the morphological annotation between the two projects. [11] These two projects are in process with new languages being added. [10, 13, 14, 17]

Computational complexity is hard to measure, and there is no standard way to quantify it yet. Ackerman and Malouf [1] propose the enumerative complexity (E-complexity) which is determined by the number of morphosyntactic distinctions in the language and how the language encodes them, and integrative complexity (I-complexity) which measures how difficult the morphological system of the language is for users. Cotterell et al. [6] verify that there is a tradeoff between the paradigm size, i.e. what we call average table size, and irregularity. The morphological complexity measurements in our project use quantitative data including the number of paradigms, the number of forms, the average inflection table size and the average sentence length, which is most similar to the E-complexity of all the dimensions proposed above.

Low-resource languages and data augmentation is one of the research topics and main concerns in computational morphology. There are also a lot of tasks and publications involving this topic, for example [2, 4, 5] etc.

3 Project description and justification

Considering the lack of integrated resource where users can find data information in UM and UD, two similar and complementary data resources for computational morphology, our project aims at filling in this gap. Therefore, our website has a *Home* page where which provide general information about the two resources, a *Language Details* page which displays and visualizes the information for each language, and a *Data Availability* page where users can learn about what languages each project has annotated or is annotation and whether each language is covered or being annotated or not in UM or UD.

3.1 Home page

The *Home* page starts with an introduction to background knowledge as well as a summary of what can be found on this website. This is supposed to provide users a quick review of the knowledge they need to read this website and overview of website information.

3.1.1 Language map

Languages are or were used by people, usually in some areas. We think that the geographical distribution of the language can provide people a more direct and concrete view of where the labeled data are from. In addition, one main concern in computational morphology or linguistic study in general now is low-resourced languages. A geographical representation of the data can provide people with a direct view of where are the languages we have data, and draw people's attention to where we have limited data.

Therefore, we made a language map with a word cloud in Fig. 3. In the language map, we use the country as a unit and when you mouse over the country, the languages used in the country which are included in UM or UD will show up as a word cloud. The advantage of this design over plotting language as dots or bubbles as WALS[1] does is that it deals with uncertainty better: though not perfect, it can reflect better the fact that languages are usually mainly used in areas not one specific clearly defined position and it is hard to draw boundaries. The interactive language map filled by the different colors relying on the counts of languages in those countries, which can provide an easy understanding of the language diversity degree of those countries.

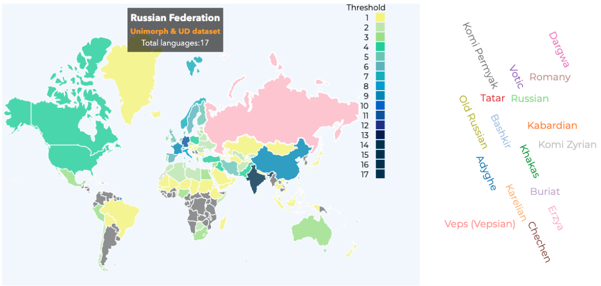


Figure 3: The interactive language map and the word cloud

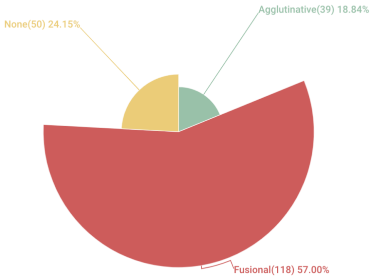


Figure 4: Pie Chart: Distribution of Typology

3.1.2 Morphological complexity

Morphological typology classifies languages based on their common morphological structures. Different typology is supposed to have a morphological system with different characteristics and languages of the same typology are supposed to share similarities in morphology. For example, agglutinative languages usually express one morphosyntactic meaning with one affix while fusional languages usually incorporate multiple morphosyntactic meanings into one affix. Therefore, typology can indicate what morphological complexity the language might have from a linguistic typological study perspective. Therefore, the distribution of languages as to typology is interesting for computational morphology and we visualized this information with an irregular pie chart (Fig. 4). In addition, the average paradigm size is a reflection of morphological complexity and we visualized the average paradigm size for each language grouped by morphology with a radial map (Fig. 5).

Language families group languages based on their genealogy. Languages of the same family are supposed to be descendants of one common ancestral language. Therefore, they are historically related and are expected to share similarities in their morphological systems, which might include aspects of morphological complexity. To visualize this distribution of languages as to language family, we used a doughnut chart (Fig. 6).

Typology and language families are two different dimensions to group languages. The genealogical relationship may be related to typological similarities and differences. To provide users a view of this information, we created a bubble chart(Fig. 7) of typologies in language families to visualize this hierarchical information.

The two scatter plots, one for the number of forms and the number of paradigms (Fig. 8) and the other for the average table size and the average sentence length (Fig. 9), with color in both indicating typology, are visualization for users to explore the hypothesis based on linguistic studies. The form-paradigm scatter plot is for exploring pattern in morphological complexity as to table sizes, which is patterns of morphological complexity out of context, and the table size-sentence length scatter plot is for exploring pattern in the relationship between morphological complexity and syntax, which is

[1] https://wals.info/

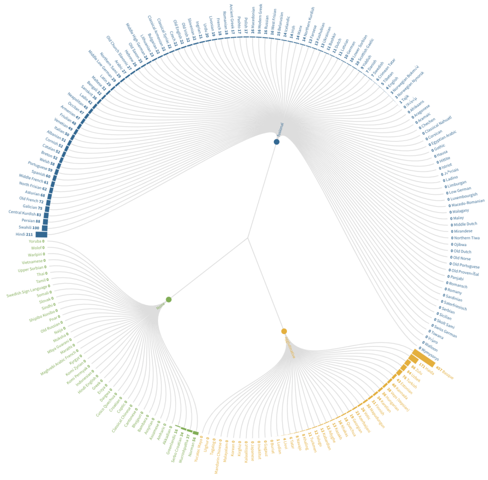


Figure 5: Radial Chart: Typology and Mean Table Size

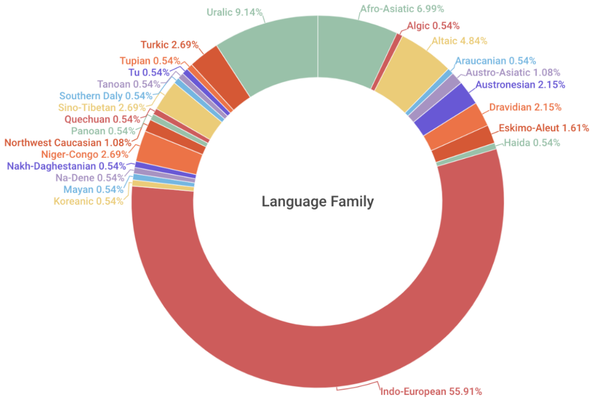


Figure 6: doughnut Chart: Language Family

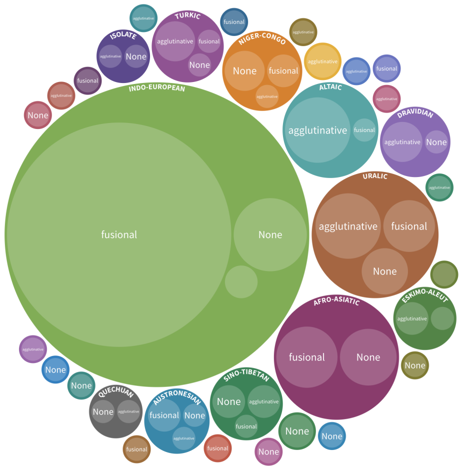


Figure 7: Bubble Chart: Language Family & Typology

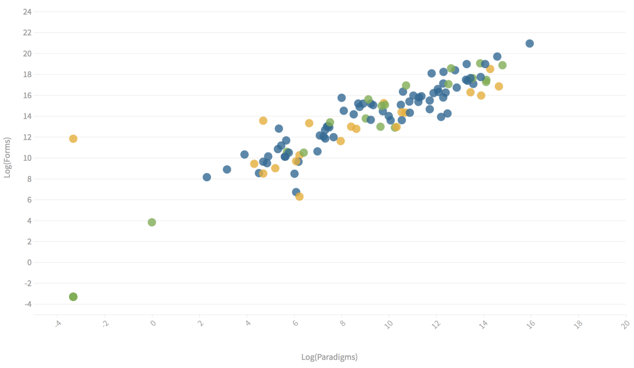


Figure 8: Scatter Plot: Forms & Paradigms

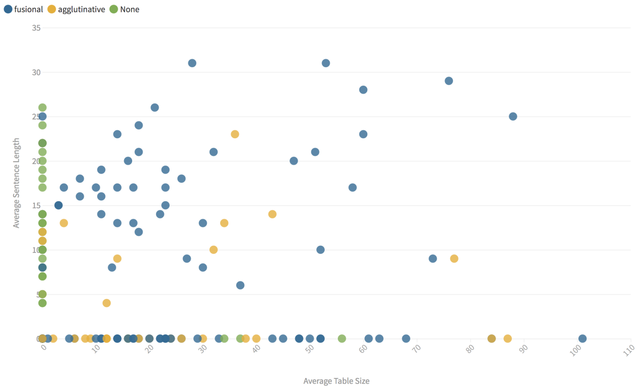


Figure 9: Scatter Plot: Mean Sentence Length & Mean Table Size

patterns of morphological complexity within context. Since linguistic studies motivate the hypothesis of certain patterns in these two aspects, whether the patterns actually exist in UM and UD data is questionable. Scatter plots are good for exploring relations between different dimensions of data.

3.1.3 Low-resource status

Though low-resourced languages are a big concern for computational morphology, this information in our visualization is categorical and straightforward. We used a treemap to group languages as to their low-resource status, which is further grouped by their language families (Fig. 10). We think this is a good design for this information because the size of the cells in the treemap conveys the number of languages in each group and the number of sub-cells in each cell indicates the number of language families or languages that are low-resources. This can provide users a quick and direct overview of low-resource languages in UM.

3.1.4 Link to Language Details

After an overview of UM and UD, it's very likely that users want to learn more about each language. Considering this, we added an example for information about one language and a link to the *Language Details* page.

3.2 Language Details page

On this page (Fig.11), users can choose the language they are interested in from a drop-down list. We used a drop-down list in order to keep a succinct layout for 199 languages. When one language is selected, the basic information about the language and its data will be listed. A pie chart visualizes the number of paradigms each

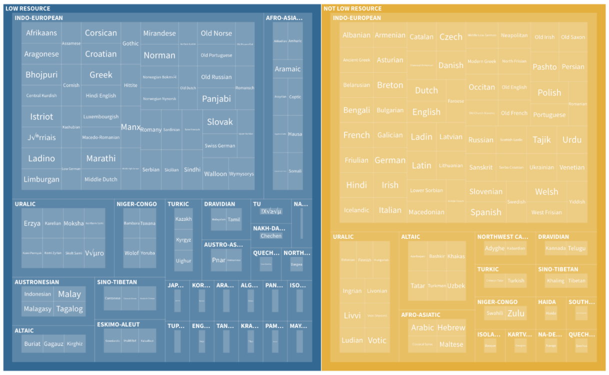
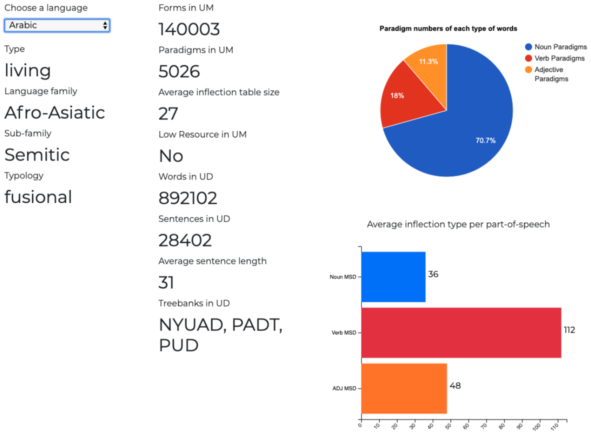


Figure 10: Treemap:Low Resource Languages



Choose a language

Arabic

Type

living

Language family

Afro-Asiatic

Sub-family

Semitic

Typology

fusional

Forms in UM

140003

Paradigms in UM

5026

Average inflection table size

27

Low Resource in UM

No

Words in UD

892102

Sentences in UD

28402

Average sentence length

31

Treebanks in UD

NYUAD, PADT, PUD

Figure 11: The interactive language detail information

part-of-speech (POS) has in the language and a bar chart visualizes the number of MSDs for each POS. This list of information is to give users an overview of the language and available data in UM and UD. The two visualizations are to provide more detailed morphological information as to POS in the data. The detailed information is also interesting to users because different POS usually have different morphological characteristics. The two charts used different libraries of JavaScript (D3 and Google chart), but we made the colors with consistency for easy understanding. We also added an animation to all numbers to attract readers' attention and reduce the boring of reading so many numbers.

3.3 Data Availability page

This page is where users can find what languages are available or in the process of being annotated in UM or UD. They can also view by language: When the mouse over the bar before the language name, its corresponding data resource will be enlarged. A sankey diagram is used for this purpose to allow the interaction from the data resource side as well as from the language side and to visualize the connections between the two sides.

4 Discussion

The visualizations and data analysis contribute to our understanding of UM and UD datasets. In particular, we want to point out the following findings:

First, most of the languages in UM are not in UD. Specially, of the 162 languages in UM or in the process of being annotated with UM schema, 101 are not in UD or being annotated with UD schema. The coverage by UM for UD data is better. Of the 106 languages in UD or being annotated with UD schema, 46 are not in UM. This discrepancy is a little surprising but makes sense considering that UM is a newer project than UD and that UD annotation is more complex than UM including grammatical information in addition to morphology. UM may have tried to cover languages UD and it's easier to use UD data to expand UM data than the other way around.

Second, countries where no language has been covered by UM or UD are mainly in Africa and South America, as is shown in the Language map in Figure 3.

Third, only agglutinative and fusional are used for UM morphological typology. The use of "agglutinative" and "fusional" are not the same as their usual linguistic definition. For example, usually English and Mandarin Chinese are classified as analytical while UM classified English as fusional and Mandarin Chinese as agglutinative. We didn't find a definition for UM uses of the two terms. For languages covered in UM and UD, Figure 4 shows that most of them (57%) are fusional languages, 18.86% are agglutinative and the others are languages in UD and UM for which we don't have typological information available.

Fourth, the average table size can vary a lot between languages. It can be as large as 457 for Basque, or as small as less than five like for English, Norweigian, Tajik Livvi and Ludian, for which Figure 5 is a visualization where users can find more information about this aspect.

Fifth, the language family that has the largest number of languages covered in the two projects is Indo-European, followed by Uralic, Afrio-Asiatic, Altaic, Turkic, etc. (See Figure 6)

Sixth, we do see one dominant typology for each language family, which is as expect about language similarity due to genealogical relationships. This pattern is observed in Figure 7.

Seventh, we didn't find any obvious pattern in the relationship between the number of paradigms and the number of forms in Figure 8, but we did find support in Figure 9 for the linguistic hypothesis about the tradeoff between morphology and syntax, i.e. languages with more inflectional forms for each dictionary form tend to have shorter sentences.

Eighth, as to low-resource status, there are much more language families which have languages in low-resource status, as is shown in Figure 10. This indicates the seriousness of the low-resourced language problem.

5 Conclusion

Our project created a website which integrates data information for languages in UM and UD, two commonly used datasets for computational morphology. This website is supposed to be a convenient reference for people who want to learn more about these two resources. The visualizations and texts convey such information and help users to explore more about the data, especially concerning morphological complexity as to typology, language family and data statistical information in the resources and low-resource status of the language.

References

[1] F. Ackerman and R. Malouf. Morphological organization: The low conditional entropy conjecture. *Language*, pp. 429–464, 2013.

[2] T. Bergmanis and S. Goldwater. Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1391–1400. Association for Computational Linguistics, New Orleans, Louisiana, June 2018. doi: 10.18653/v1/N18-1126

[3] T. Bergmanis and S. Goldwater. Training Data Augmentation for Context-Sensitive Neural Lemmatizer Using Inflection Tables and Raw Text. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4119–4128. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. doi: 10.18653/v1/N19-1418

[4] T. Bergmanis, K. Kann, H. Schütze, and S. Goldwater. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pp. 31–39. Association for Computational Linguistics, Vancouver, Aug. 2017. doi: 10.18653/v1/K17-2002

[5] J. Buys and J. A. Botha. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1954–1964. Association for Computational Linguistics, Berlin, Germany, Aug. 2016. doi: 10.18653/v1/P16-1184

[6] R. Cotterell, C. Kirov, M. Hulden, and J. Eisner. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342, Mar. 2019. doi: 10.1162/tacl_a_00271

[7] R. Cotterell, C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. Mielke, G. Nicolai, M. Silfverberg, D. Yarowsky, J. Eisner, and M. Hulden. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pp. 1–27. Association for Computational Linguistics, Brussels, Oct. 2018. doi: 10.18653/v1/K18-3001

[8] R. Cotterell, C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, P. Xia, M. Faruqui, S. Kübler, D. Yarowsky, J. Eisner, and M. Hulden. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pp. 1–30. Association for Computational Linguistics, Vancouver, Aug. 2017. doi: 10.18653/v1/K17-2001

[9] R. Cotterell, C. Kirov, J. Sylak-Glassman, D. Yarowsky, J. Eisner, and M. Hulden. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 10–22. Association for Computational Linguistics, Berlin, Germany, Aug. 2016. doi: 10.18653/v1/W16-2002

[10] K. Dobrovoljc, T. Erjavec, and S. Krek. The universal dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pp. 33–38. Association for Computational Linguistics, Valencia, Spain, Apr. 2017. doi: 10.18653/v1/W17-1406

[11] A. D. McCarthy, M. Silfverberg, R. Cotterell, M. Hulden, and D. Yarowsky. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 91–101. Association for Computational Linguistics, Brussels, Belgium, Nov. 2018. doi: 10.18653/v1/W18-6011

[12] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666. European Language Resources Association (ELRA), Portorož, Slovenia, May 2016.

[13] J. Nivre and C.-T. Fang. Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May, Gothenburg Sweden*, number 135, pp. 86–95. Linköping University Electronic Press, 2017.

[14] S. Pyysalo, J. Kanerva, A. Missilä, V. Laippala, and F. Ginter. Universal dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pp. 163–172. Linköping University Electronic Press, Sweden, Vilnius, Lithuania, May 2015.

[15] M. Silfverberg and M. Hulden. Automatic morpheme segmentation and labeling in universal dependencies resources. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pp. 140–145. Association for Computational Linguistics, Gothenburg, Sweden, May 2017.

[16] J. Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*, 2016.

[17] D. Taji, N. Habash, and D. Zeman. Universal dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pp. 166–176. Association for Computational Linguistics, Valencia, Spain, Apr. 2017. doi: 10.18653/v1/W17-1320