Visualizing Inferred Morphotactic Systems

Haley Lepp
University of Washington
Department of Linguistics
hlepp@uw.edu

Olga Zamaraeva
University of Washington
Department of Linguistics
olzama@uw.edu

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Abstract

We present a web-based system that facilitates the exploration of complex morphological patterns found in morphologically rich languages. The need for better understanding of such patterns is urgent for linguistics and important for cross-linguistically applicable natural language processing. We give an overview of the system architecture and describe a sample case study on Abui [abz], a Trans-New Guinea language spoken in Indonesia.

1 Introduction

Understanding and describing morphological patterns is a fundamental task in both documentary linguistics and the development of language technology. Many low-resource or underdescribed languages evince a high degree of morphological complexity, with large numbers of distinct affix types and many affix tokens possible within a single word. At the same time, building language technology for morphologically complex low-resource languages requires a rule-based morphological analyzer when datasets are not large enough for ML approaches (see Garrette et al. 2013; Erdmann and Habash 2018, *inter alia*). Our contribution is within the context of the AGGREGATION project, which aims to automatically infer broad typological characteristics and morphological patterns for understudied languages (Bender et al., 2013; Zamaraeva et al., 2017). We developed this visualization tool to help linguists to understand the morphological system implicit in large datasets and to refine automatically generated grammar specifications which model that morphological system. Thus we expect this tool to directly assist in language description. Because linguistic typology depends on accurate language description, and truly language-independent NLP depends on linguistic typology (see Bender 2011 and Gerz et al. 2018), we also anticipate long-term benefits for NLP.

We present a visualization component for the MOM morphological inference system (see §2.1) which takes as input a collection of morpheme-segmented, glossed text and produces a set of hypotheses about classes of stems and affixes, and the cooccurrence and ordering possibilities between them. This set of hypotheses is cast as a grammar specification that can be used by the Grammar Matrix customization system (Bender et al., 2010) to automatically create an implemented grammar capable of morphological parsing. Our system helps the linguist visualize and explore these hypotheses, facilitating both further linguistic theorizing of the dataset and the production of a more accurate implemented grammar. The grammar can be used to produce annotations for additional unglossed data, as in Zamaraeva et al. 2017, because the inferred system generalizes beyond the specific combinations of morphemes observed.[1]

2 System overview

2.1 Back-end

The morphological graph that our system visualizes comes from the MOM morphological inference software (Wax, 2014; Zamaraeva, 2016). MOM outputs a directed acyclic graph specified in DOT (Gansner et al., 1993) where nodes are inflectional classes of words and position classes of affixes, and edges reflect the ordering possibilities of those affixes. This graph is translated by the Grammar Matrix customization system to

[1] The (frequently updated) Alpha-version of the system can be accessed via http://uakari.ling.washington.edu/aggregation/. The demonstration video is at https://youtu.be/qn96Zg-6wkE. All of the code and sample data are available in the repository: https://git.ling.washington.edu/agg/mom

Proceedings of NAACL-HLT 2019: Demonstrations, pages 127–131
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

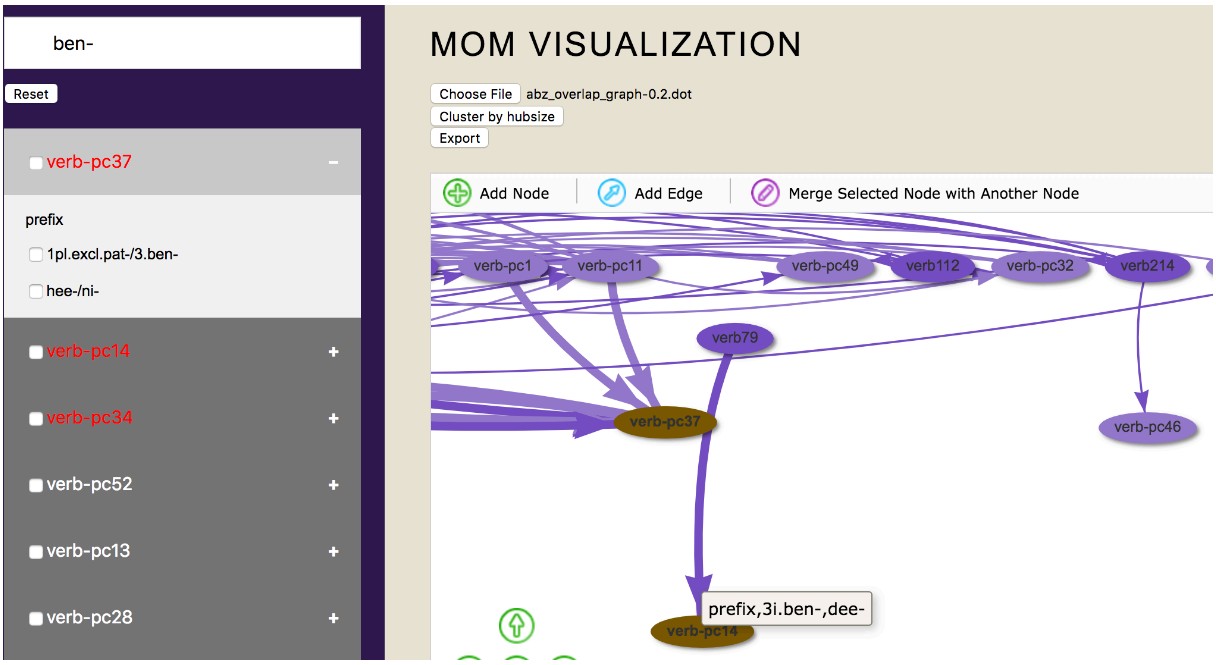

Figure 1: A portion of the visualization window, with two of the nodes selected. The contents of the nodes can also be explored via the sidebar which has a search field. Moving nodes around helps the user explore the denser parts of the graph.

a grammar that includes: (i) lexical types, each instantiated by (ii) lexical entries, and (iii) lexical rule types, defining position classes, and each instantiated by (iv) lexical rules, which each attach a specific affix. The position classes also define the possible order of attachment of morphemes, known as the morphotactics of the language.

These graphs tend to be quite large, because of the underlying complexity of the systems they describe and because of noise introduced by glossing inconsistencies and the inference process. For example, the graph we use in our case study (see §3) is the visualization of an analysis of a dataset consisting of 8609 glossed verb tokens and includes 65 nodes (20 for stems and 45 for affixes).

2.2 Visualization window

The first component of our visualization front-end, built using the Network library of vis.js,[2] is a visual network of all the morphological data that MOM outputs as a text (DOT) file. When a user imports a DOT file into the system, the gold, right-hand side of the webpage will load a visual representation of the graph.

Roots are represented as dark purple nodes, while prefixes are a medium purple, and suffixes are light purple. The morphemes are labeled with a unique ID, such as "verb1", and hovering over a node with a cursor will bring up a truncated list of root and affix spellings assigned to that stem or position class in the inferred graph.

The relationships between these morphemes are represented by directed arrows, showing in what ways the roots and affixes in a language can be combined. The root will point to the first affix, such as a prefix, and that affix will point to whichever next affix is possible.[3]

There are a number of actions that the user can take within this visualization. When the user clicks on a node, the node and all its associated linkages become bold, to aid the user in viewing the relationships. The user can also drag nodes and arrows to examine particular linkages, and zoom in on specific relationships. A "cluster" button above the visualization will combine nodes with many linkages into one, allowing the user to better see the morphemes with fewer connections. Double-clicking on the cluster will reset the graph to a full view.

[2] http://visjs.org/docs/network/

[3] On the analysis provided by the backend system, word construction starts with the root, prefixes are added from the root out, and suffixes are attached last, again from the root out. Thus the leftmost prefix is the input for the first suffix.

2.3 Manipulation

The primary purpose of giving researchers a visual representation of a graph is to give a new perspective that may elicit the discovery of new patterns or possibly inconsistencies in the glossing of the underlying data. A secondary purpose is to help the user make improvements to the graph, so that the resulting grammar output by the Grammar Matrix customization system will also function better. For both purposes, it is important that the user be able to manipulate the analysis while looking at the visual representation. Thus the second component of the interface is the functionality for the user to manipulate the graph within the visualization. At the top of the network window, there is a purple button called "Edit". When clicked, the Add Edge button appears, with Remove Edge appearing once an edge is selected.

The simplest modification that the user is expected to make to the graph is adding and deleting edges. MOM typically infers lots of possibilities for morpheme orderings from the data (see Figure 1). Some of these orderings may reflect noise in either the underlying data or in the inference results; there may also be missing orderings that the linguist is aware of based on their expertise. Either way, the morphological hypothesis presented by the system will normally require revisions. To add an ordering possibility between a stem and an affix or between two affixes, the user may add a directed arrow by clicking on a node and dragging the new arrow to another node before releasing the click. If the user notices an incorrect linkage between the morphemes, they can select the arrow and click "remove edge".

When the user clicks on a node, two more options appear: Remove Edge (explained above) and Merge Nodes. "Merge" allows the user to combine two previously separate nodes. This is useful if the researcher notices that the MOM inference system has done insufficient generalization and failed to combine two sets of affixes into one position class or two sets of stems in one inflectional class. "Split", when implemented, will split a node into two if the visualization allows the researcher to see that the graph incorrectly combines two position or inflectional classes. With "Split", we will be adding the functionality to select subsets of node contents (e.g. some but not all of the stem spellings associated with a node). This will also add the functionality to merge a subset of a node with another node.

Once the user has made changes to the graph, they can click the "Export" button above the window. The system will save the adjusted analysis to a downloadable DOT file, which can be imported into MOM or reloaded into the visualization system at a later time.

2.4 Sidebar

The third component of the interface is the purple sidebar on the left-hand side of the page. When the user loads a graph, the sidebar populates with a list of every node in the graph. When the user clicks on a plus sign on a row in the list, the row will expand, listing every associated content item, such as stems for inflectional classes and affix spellings and glosses for position classes.

In a large graph, it can be difficult to find a specific morpheme in the visual representation. At the top of the sidebar, there is a search box in which the user can search for any node ID or spelling, or gloss. The user can reset the results by clicking the "Reset" button.

3 Case study

In this section we demonstrate how the system can be used to refine morphological hypotheses suggested automatically by the MOM system and in the process discover a glossing inconsistency in the underlying data.

We run the MOM system (the back end) on the Abui dataset (Kratochvíl, 2017, abz; Trans-New Guinea). The dataset comes from a multi-year fieldwork project which involves data collection, transcription, and linguistic analysis, including that of morphology. Our tool targets this last stage. Example 1 illustrates the original input.

(1) Na aloba he-mia
1SG.AGT thorn.LOC 3UND.LOC-take.IPFV
'I am taking out the thorn.' [abz; N12.064]

Glossed examples like this one present an analysis of morphology on one particular sentence; the linguist will want to generalize to a set of hypotheses about the general morphological system in the language. For example, a question might be: Which prefixes occur with which verb stems (Zamaraeva et al., 2017) and is there any semantic coherence to the verb inflectional classes identified this way (Kratochvíl and Delpada, 2015)? In order to answer this kind of question more fully, linguists

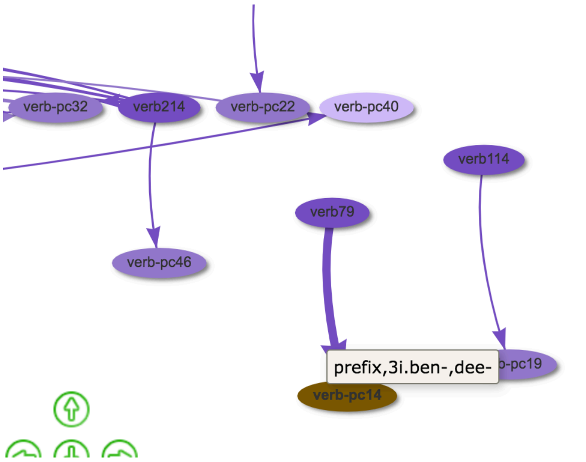

Figure 2: Some nodes ended up separate from the main portion of the morphological graph.

may find it helpful not only to find all possible cooccurrences of, say, prefixes and verbs, but also to visualize them as a graph. The MOM system produces a DOT file which represents all possible cooccurrences, and our tool visualizes it. The morphological hypotheses that the DOT file represents are the result of all of the cooccurrence patterns found in the original data being compressed in such a way that nodes which share more than 20% of edges are combined (Wax, 2014).[4]

Examining the graph, we notice that, while most of it is connected, there are a couple of nodes on the right that stand alone. Upon inspection, it turns out that the node "verb79" contains only one stem, the serial verb *l* ('give'). The position class that it is linked to, "verb-pc14", contains prefixes which are glossed as 3rd person, benefactive. After we search for "ben-", we see that there are two more nodes in the graph which contain affixes with this gloss (see Figure 1). Now we can hypothesize that two or all three of them actually should be one position class and we can merge them in the graph as described above and the result shown in Figure 3. Furthermore, after we search for "give", we discover that the serial verb *l* is present in the graph both as a root and as a prefix (*l*-). In other words, the visualization helped us to discover that a single morpheme (pairing of form and meaning) was analyzed in two different ways in the annotations (as a stem and as a prefix). This could be the result of simple inconsistency in glossing in the dataset or it could be indicative of variation in the language meriting further attention by the linguist. Finally, recall that the graph is a specification of a morphological grammar, so any revisions of the hypotheses that we make in this fashion can be tested by creating a grammar automatically using the Grammar Matrix customization system and then parsing held-out data.

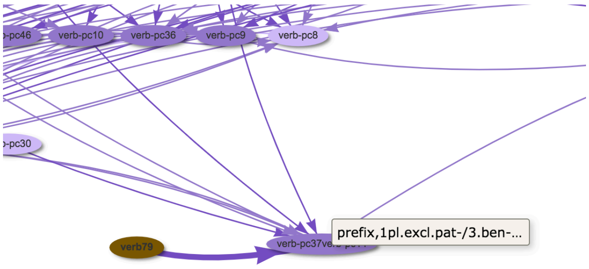

Figure 3: Two nodes for the benefactive prefix were merged in one.

4 Future work

Among the technical improvements we envision are more manipulation functions, such as the ability to split a node in two. We also plan to allow the user to manipulate the graph from the sidebar in addition to from the visual representation. With these improvements in place, we plan to invite field linguists to try out the system and map out further features based on user feedback.

A second direction for future work is to update the back-end MOM system so it can run inference constrained by user-specified improvements to the graph. Our goal here will be to assist the linguist in discovering the further implications of split/merge decisions and more generally facilitate collaborative human-machine discovery of morphological systems.

5 Conclusion

We have presented a visualization tool that allows users to explore automatically inferred morphological grammar specifications based on linguistically annotated datasets. For the linguist-user, this tool has two purposes: (i) to help them better understand the systems inherent in their datasets and annotations and (ii) to help them refine the resulting morphological analyzer so as to produce better annotations of unglossed data. In the longer term, more thorough glossing of linguistic datasets and analyses of morphological systems benefits both linguistic typology and, ultimately, NLP.

[4] The compression rate is a configurable parameter.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1561833. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We thank Swetha Ramaswamy for developing the prototype for our system; Russel Hugo and David Inman for participating in the prototype interviews; Tara Clark for contributing to the initial UI implementation discussion; and František Kratochvíl for motivating discussion and for sharing the Abui data.

References

Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Emily M Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8:1–50. ISSN 1570-7075.

Emily M Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. August 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 74–83, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2710.

Alexander Erdmann and Nizar Habash. 2018. Complementary strategies for low resourced morphological modeling. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 54–65.

Emden R. Gansner, Eleftherios Koutsofios, Stephen C. North, and Kiem phong Vo. 1993. A technique for drawing directed graphs. *IEEE transactions on software engineering*, 19(3):214–230.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 583–592.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327.

František Kratochvíl. 2017. Abui Corpus. Electronic Database: 162,000 words of natural speech, and 37,500 words of elicited material (February 2017). Nanyang Technological University, Singapore.

František Kratochvíl and Benidiktus Delpada. 2015. Degrees of affectedness and verbal prefixation in Abui (Papuan). In Stefan Müller, editor, *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, pages 216–233, Stanford, CA, 2015. CSLI Publications.

David Wax. 2014. Automated grammar engineering for verbal morphology. Master's thesis, University of Washington.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150.

Olga Zamaraeva, František Kratochvíl, Emily M Bender, Fei Xia, and Kristen Howell. 2017. Computational support for finding word classes: A case study of Abui. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 130–140.