Feature Comparison across Typological Resources

Tifa de Almeida, **Youyun Zhang**, **Kristen Howell**, and **Emily M. Bender**

Department of Linguistics, University of Washington, Seattle, WA, USA

{trda, youyunzh, kphowell, ebender}@uw.edu

1 Introduction

We explore in this abstract the relationship between a typological database (WALS) and a grammar specification system (the LinGO Grammar Matrix).

The LinGO Grammar Matrix (GM) customization system automatically generates a custom HPSG grammar based on user inputs to a web-based questionnaire (Bender et al., 2002, 2010). One of its goals is to make the process of creating new grammars easier for linguists of all backgrounds working on testing linguistic hypotheses.

The World Atlas of Linguistics Structures (WALS), often considered a typological atlas due to its detailed geographical data, distinguishes itself from other typological databases such as Terraling (Terraling) by the quantity and the quality of data it contains. WALS was built upon the work of 55 authors who classified over 2,500 languages according to 192 features,[1] and integrated this information with geographical coordinates for each language. It has been used to discover universal typological implications (Daume III and Campbell, 2007), compare phylogenetic relationships within feature-based language clusters (Georgi et al., 2010) and provide a benchmark for automatic typological feature identification from corpora (Lewis and Xia, 2008).

We explore this database in an effort to assist the creation of custom grammars for the GM user. We develope a method to determine what is the overlap between the GM questionnaire and WALS features and how they correspond to each other. In the following sections, we detail how WALS features can be mapped to the Grammar Matrix and what conclusions can be drawn from these mappings. To illustrate this process we provide examples and conclude with a discussion of the potential impacts of these matches and how they may be applied in future work.

2 Methodology

There is a fundamental difference between the features defined in the WALS database and the ones elicited by the GM user interface: While both are built with features extracted from detailed grammars and current typological literature, WALS is fundamentally a reference database of language typology. Accordingly, WALS features classify typological information about a language but are generally not concerned with all the detail that would be required to implement language-specific grammars (e.g. the particular form affixes take).[2]

Due to this, we developed a simple method to determine which WALS features match which GM features to determine to what extent WALS features can be imported and utilized in the grammar customization process. After studying the documentation for a GM feature, for example Adnominal Possession, we examine WALS' inventory of features looking for key terms in titles that correspond to the GM phenomenon, such as Feature 24A Locus of Marking in Possessive Noun Phrases (Nichols and Bickel, 2013). We assess the values of the feature and organize them with the corresponding questionnaire item. An example of this 1-to-1 pairing can be seen in Table 1.

This is where a careful interpretation of the documentation for each system is necessary. WALS utilizes the term locus to refer to a head-dependent marking relationship (Comrie, 2013), designating the possessed noun as the head noun and the possessor as the dependent. The GM refers to the

[1] This dataset is sparse however: different features are specified for different languages.

[2] WALS also contains significant information about linguistic properties outside the morphological, syntactic and semantic information required by the GM, including phonological and lexical features.

Grammar Matrix	WALS	
Morpheme Placement	**Feature 24A**	**Number of Languages**
On the possessum	Possessor is head-marked	78
On the possessor	Possessor is dependent-marked	98
On both the possessor and the possessum	Possessor is double-marked	22
No possessive morphemes appear	Possessor has no marking	32
—	Other types	6

Table 1: Correlation of GM Adnominal Possession morpheme placement options and WALS Feature 24A Locus of Marking in Possessive Noun Phrases. Data available for 236 languages.

possessor and the possessum (Nielsen and Bender, 2018). Therefore we pair the values "on the possessum" (GM) with "possessor is head-marked" (WALS). This analysis is repeated with all values for each feature.

Another point of consideration this particular feature brings up is that after asking about the position of the possessive morpheme, the GM posits further clarifying questions that depend on which option was chosen. If the morpheme appears "on the possessor", the user is asked if it is an affix, separate word or clitic. When the option "on the possessor" is chosen, the GM offers the user the option to add a feature constraint, such as Case. This kind of information, along with the orthography and distribution of the morphemes, is regrettably not provided by WALS and must remain under the user's purview to add.

Below we offer two more examples of WALS/GM features to illustrate the challenges of feature mapping when a one-to-one correspondence cannot be found or is not particularly useful.

2.1 Case Marking Strategy

The core case marking section of the GM (Drellishak, 2008) corresponds to WALS Feature 98A (Comrie, 2013), entitled *Alignment of Case Marking of Full Noun Phrase*. The latter designates the argument abbreviations according to Dixon (1994) – A (agent of a transitive verb), O (object of a transitive verb) and S (subject of an intransitive verb) – whereas WALS uses P (patient) instead of O (Comrie, 1978).

The GM questionnaire gives the user the option to check which case marking strategy is being used and also what each case is called, e.g. ergative. By extracting the values of feature 98A from WALS, the user is given a head-start at this. For the 190 languages specified for this feature in WALS, the GM user has a readily usable source of information for the first of these steps.

2.2 Number of Cases

WALS feature 49A (Iggesen, 2013), defined for 261 languages, maps how many cases a language contains. One would think that knowing the number of cases would be helpful in building a grammar, but it is actually not. Without also knowing what each case is called, this feature could only notify the user that they must manually add N cases, which is not useful for our purposes.

3 Conclusion and Future Work

Having reviewed 33 WALS features, we estimate that about 20 of them (10.4% of the total) can be usefully imported into the GM system to facilitate the grammar generation process for the user. This corresponds to about 8.5% of the GM's grammar specification options.

Our work identifying which features can be mapped and in what way could support an API that extracts the pertinent information from WALS when the user starts a custom grammar. The first page of the user interface asks the user to input the language ISO code, which is also used in the WALS database. The user would be given a choice of importing this information or not, and should they choose to do so would be shown a notification detailing how many features were found.

Additionally, our mapping of WALS to GM features could support work on automatically answering the GM questionnaire. The AGGREGATION project (e.g. Zamaraeva et al., 2019) presently approaches this problem by inferring GM grammar specifications from collections of interlinear glossed text. Where WALS features map to GM features, and WALS values are available for the language at hand, the WALS values can potentially be used to guide grammar specification inference, as explored in Zhang et al. 2019.

References

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Res. Lang. Comput.*, 8(1):23–72.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02: Grammar Engineering and Evaluation*.

Bernard Comrie. 1978. Ergativity. In Winfred P. Lehmann, editor, *Syntactic Typology: Studies in the Phenomenology of Language*, pages 329–394. University of Texas Press, Austin.

Bernard Comrie. 2013. Alignment of case marking of full noun phrases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hal Daume III and Lyle Campbell. 2007. A bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Robert M. W. Dixon. 1994. *Ergativity*. Cambridge University Press, Cambridge.

Scott Drellishak. 2008. Complex case phenomena in the grammar matrix. In *Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar, National Institute of Information and Communications Technology, Keihanna*, pages 67–86, Stanford, CA. CSLI Publications.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 385–393, Beijing, China. Coling 2010 Organizing Committee.

Oliver A. Iggesen. 2013. Number of cases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Johanna Nichols and Balthasar Bickel. 2013. Locus of marking in possessive noun phrases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Elizabeth Nielsen and Emily M. Bender. 2018. Modeling adnominal possession in multilingual grammar engineering. In *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar, University of Tokyo*, pages 140–153, Stanford, CA. CSLI Publications.

Terraling. Available online: http://test.terraling.com. Accessed: 26 April, 2019. [link].

Olga Zamaraeva, Kristen Howell, and Emily M. Bender. 2019. Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, volume 1 Papers, pages 28–38, Honolulu, Hawai'i.

Youyun Zhang, Tifa de Almeida, Kristen Howell, and Emily M. Bender. 2019. Using typological information in wals to improve grammar inference. Unpublished paper, submitted to TypNLP.