

NLP Project Summary

Diana Maaßen (12345096) Philip Liszt (01227444)
Teodor Esaiasson (12443590) Veneta Grigorova

January 2026

1 Task Description

In this project, the aim was to investigate the linguistic rules known as Greenbergs universals Greenberg (1963), using large-scale language data. At the start of the project we selected a subset of 16 rules for further analysis.

While the original task was focused around the Greenberg Universals, we have instead worked with a set of linguistic universals that were inspired by them, while not being exactly the same. Five of the 16 rules (Rule 3, 4, 5, 18 and 41) are Greenberg universals, and the rest is in a similar style but not actual Greenberg rules. The full list of rules investigated can be found in the "README.md" inside the submission folder.

These universals describe how features of sentence structure relate to one another, for example how the position of a verb relative to the object affects the placement of adjectives in a sentence. The availability of linguistic databases covering thousands of languages motivated us to reassess these universals on larger and more diverse empirical evidence.

By translating the rules into machine-processable criteria, we evaluated them using both global typological data and annotated language corpora. Our analysis focused on assessing the degree of empirical support for a subset of 16 universals and examining whether violations follow geographic or usage-based patterns. Furthermore, we applied simple machine learning methods that yield interpretable results to investigate whether existing universals can be derived from the data and whether new similar rules learned from it can offer improvement on the investigated rules.

2 Used Datasets

World Atlas of Language Structure (WALS) is a large typological database that stores grammatical and structural features of languages worldwide at language-level. We used it to test the Greenberg universals empirically across many lan-

guages.

Universal Dependencies (UD) is a collection of linguistically annotated sentences. We used German corpora from UD to compare our language-level analysis of the rules with evidence from individual sentences.

3 Key Challenges

A key challenge was identifying the appropriate linguistic features for each rule, as the universals are formulated in natural language a single condition could relate to multiple features. In addition, the WALS data provide general language-level descriptions, while natural languages have many exceptions that are not fully captured in the data.

During the process of translating the rules and simplifying them so that they could be mapped to language properties in the WALS data, there was a mix up with the original wording of some of the rules. This is the reason why not all rules analyzed are actual Greenberg Universals, but instead only in a similar style. While this of course lessens the connection with the original task, the analysis still contains information about general properties that holds true for larger subsets of languages and the possibility to predict properties of a language based on other properties.

Another challenge arose in the corpus-based analysis using Universal Dependencies data. Greenberg's universals are formulated at the level of entire languages, whereas UD provides syntactic annotations at the level of individual sentences. For german in particular, flexible word order and verb-second phenomena make it difficult to determine VO and OV patterns based on surface word order alone. We therefore identified verbs and their objects using UD dependency relations and classified VO and OV order by comparing their token positions within the sentence. As a result, the corpus-based evaluation should be interpreted as an operationalized approximation of the universals rather than a direct one-to-one implementation.

4 Main Findings and Limitations

The Greenberg universals that link basic word order to adposition type (Rules 3 and 4) were strongly supported. In approximately 90-95% of languages for which the precondition holds, the predicted adposition type was observed. Greenberg Rules 5 and 18, which concern adjective placement, showed weaker conditional support, with the conclusion holding in about 70-80% of the relevant languages. For Greenberg Universal Rule 41, which states that *"If in a language the verb follows both the nominal subject and nominal object as the dominant order, the language almost always has a case system."* only 56% of the languages satisfying the precondition exhibited the predicted case system.

Among the additional, non-greenberg rules evaluated, Rules 16, 17, and 26 exhibited very strong conditional support, with 95-100% of languages satisfying

the rule when the precondition was met. Rules 6, 18, and 25 showed moderate conditional support at around 70%. In contrast, Rules 19, 20, 21, and 24 displayed weak conditional support, with approximately 50% conformity given the precondition, while Rule 23 showed particularly poor conditional support (around 20%), indicating that the opposite is more common. For Rules 19, 20, 21, 23, 24, and 41, the geographic regions in which exceptions occur vary depending on the specific rule considered.

Interpretable machine learning models were able to rediscover several classic universals and also identify new linguistic rules with high predictive accuracy. To complement the typological analysis based on WALS, we selected universals using German Universal Dependencies corpus data. This allowed us to test whether the predicted patterns are reflected in real sentence usage. We analyzed Rules 20, 23, and 24 using the German HDT treebank in CoNLL-U format. For Rule 20, only 13% of 845 relevant noun phrases followed the predicted pattern, while 87% violated it, showing that German systematically contradicts this rule. Rule 23 was fully supported: in all 4,887 relevant cases, adjectives preceded the noun in VO word order. In contrast, Rule 24 was completely violated, as adjectives always preceded the noun in the 7,571 tested OV cases. The corpus-based analysis demonstrates that linguistic universals should be interpreted as tendencies rather than strict rules when tested on real language data.

The analysis was limited by the high-level language descriptions provided by WALS, which do not capture all internal variation within a language. In addition, WALS does not provide complete coverage of all languages worldwide. Finally, the selected corpus may not be fully representative of the German language as a whole.

5 Future Work

Further analysis of the geographical patterns for rule compliance and exceptions, with respect to language families, could offer opportunities to improve on the linguistic universals. One possibility could be to include the family in the machine learning approach for creating new universals, as an approach to find reaching to finding more rules with a high accuracy.

As mentioned in key challenges, WALS data provides general language descriptions that have many exceptions. The analysis of the universals could therefore be aided by the further usage of different datasets which capture more details about each language, as to deepen the understanding of their generalizability.

6 Individual Contributions

The work in the project has been evenly divided between the team members, with everyone having one main topic to focus on. Although the work was divided, the group made sure to regularly have discussion about the progress of different topics and how they could be integrated for a coherent final result. The topics were divided according to the following:

Cross tables of rule compliance and geographical analysis of exceptions: Philip Liszt, Veneta Grigorova

Rule based classifier and Decision Trees for creation of new linguistic universals based on WALS data: Teodor Esaiasson

Qualitative analysis of linguistic universals based on UD data : Diana Maaßen

References

- Greenberg, Joseph H. (1963). "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements". In: *Universals of Human Language*. Ed. by Joseph H. Greenberg. Cambridge, Mass: MIT Press, pp. 73–113.