Classifying Languages by Dependency Structure

Typologies of Delexicalized Universal Dependency Treebanks

Xinying Chen
School of International Studies
Xi'an Jiaotong University, China
Department of Czech Language
University of Ostrava, Czechia
xy@yuyanxue.net

Kim Gerdes
LPP (CNRS)
Sorbonne Nouvelle
France
kim@gerdes.fr

Background: The general philosophy of UD is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. By doing so, UD expects such a schema, as well as the treebank data, would be 'satisfactory on linguistic analysis for individual languages', meanwhile, it would also 'be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families'.

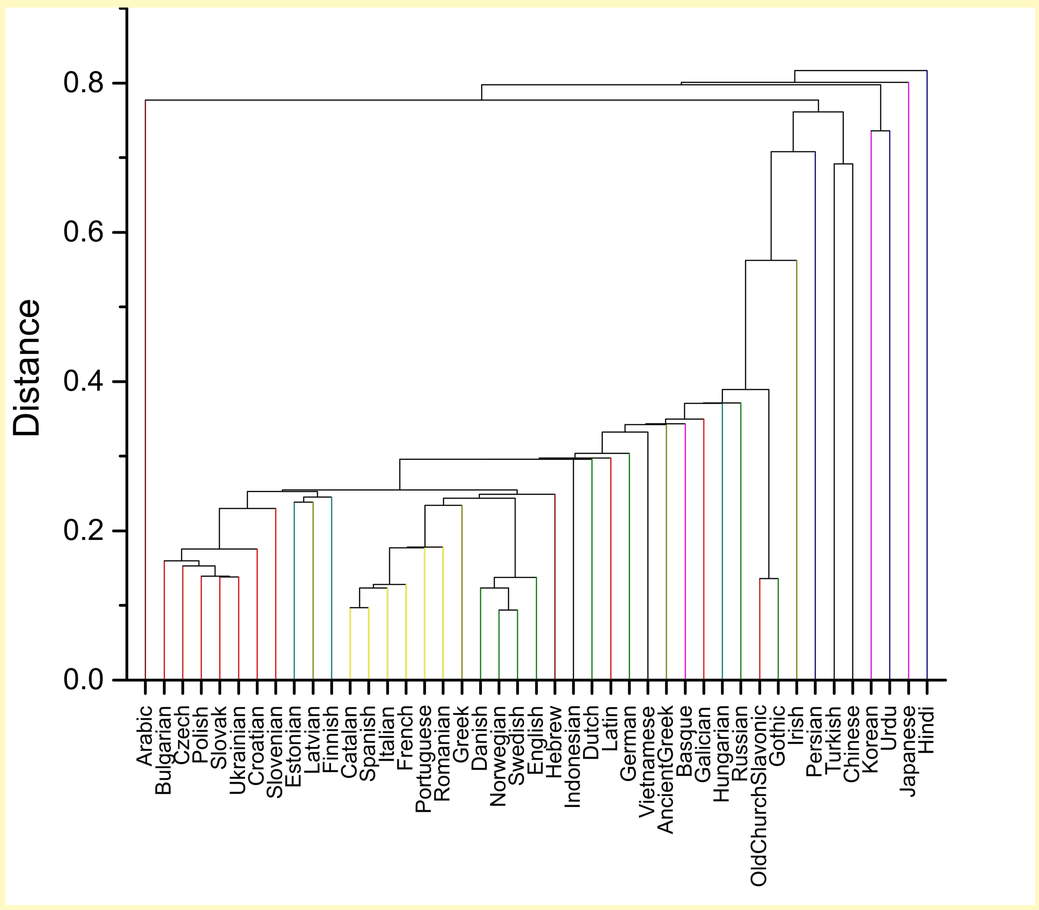
Related research: Modern language typology research (Croft 2002; Song 2001), mostly based on Greenberg (1963), generally puts much emphasis on the syntactic order (word order), in particular of the principal components in relation to their governing verb (Haspelmath et al. 2005).

Question: Can UD be used for language typology study and reveal the similarity and diversity between language families? If so, then how?

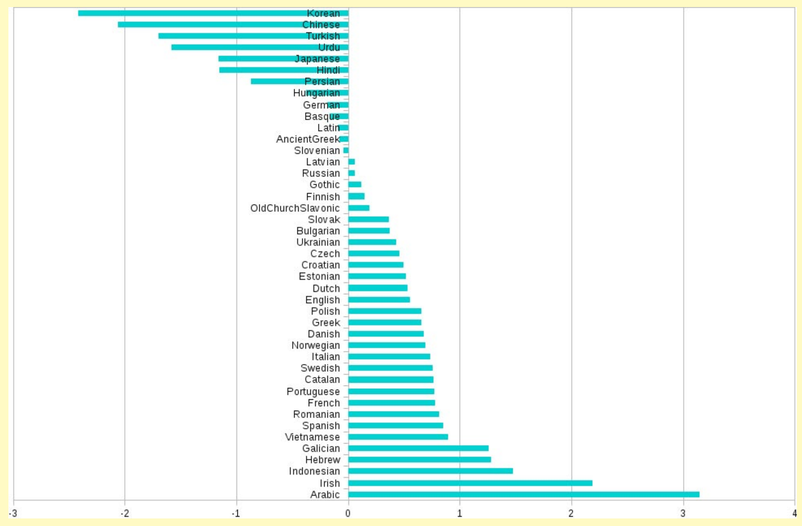
Data: 70 treebanks of 50 languages, 63 of which have more than 10,000 tokens.

What to measure: Word order + dependency distance (DDD-Directional Dependency Distance) of all dependencies (Not only SV/VS, VO/OV, AdjN/NAdj…)

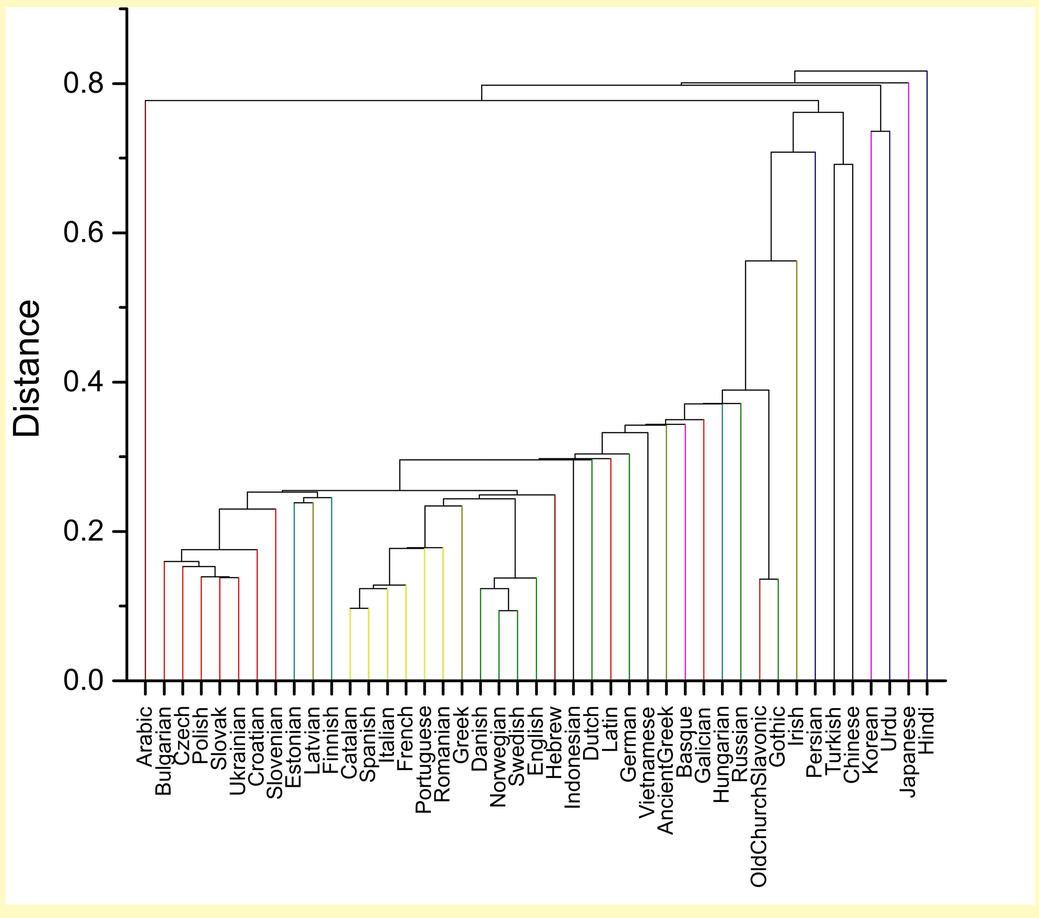
$$DDD(R) = \frac{\sum_{r \in R} distance(r)}{frequency(R)}$$



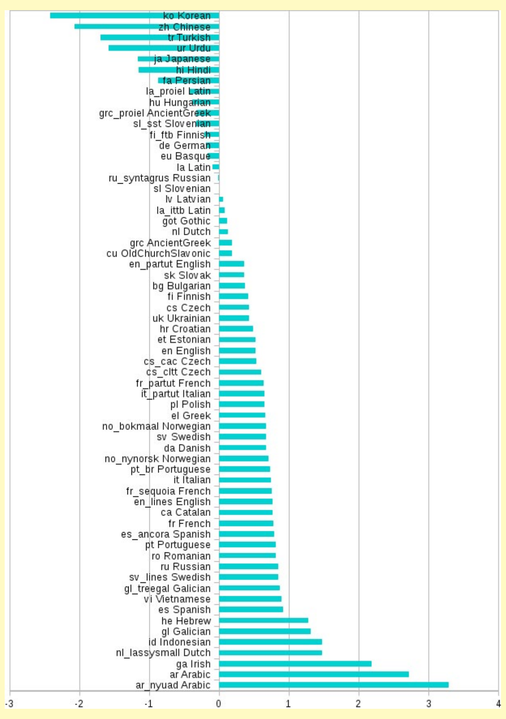
Dendrogram of distance × frequency clustering per language



Languages ordered by dependency distance



Dendrogram of distance × frequency clustering per corpus



Treebanks ordered by dependency distance
with positions for English and French

Conclusion:

1. Yes. UD really can be used for language typology study and reveal the similarity and diversity between language families.

2. We should use more detailed and quantiative measurements of word order to study language typologies.