Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks
Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

Xinying Chen
Xi’an Jiaotong University
University of Ostrava

Kim Gerdes
LPP
Sorbonne Nouvelle
Goal

• Recognizing language families based on purely empirical structural data

• Differences become
  – Quantifiable
  – Localizable

• Side effect:
  – Assessing treebank coherence and quality
Method

• Delexicalize the UD treebanks
• Compare the remaining structures
Steps

• Keep treebanks >10k tokens
• Keep only core syntagmatic relations:
  – removing fixed, flat, conj, and root
• compute
  – relative frequency distributions of dependency functions
  – Directional Dependency Distance
    \[ \text{DDD} = \text{dependency distance} \times \text{direction} \]
    \[ DDD(R) = \sum_{r \in R} \frac{\text{distance}(r)}{\text{frequency}(R)} \]
Is UD good enough?
Dendrogram of DDD vectors
PCA of DDD vectors
PCA of POS frequencies
Dendrogram of distance $\times$ relative frequency: per language
Dendrogram of distance × relative frequency: per corpus
Results

• Good measures can find language groups
  – Also on dirty data
• This makes syntactic typology
  – Empirical
    • *Number vs existence of phenomena*
  – Quantifiable
• Treebank quality assessment:
  – Is it typology?
  – Or simply an error in the annotation scheme?
• Things can only get better as UD improves
  – the quality
  – the scheme
• *Come to see our poster to discuss further!*
Grazie mille!
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’.