Autogramm: Simultaneous development of treebanks and corpus-driven grammars

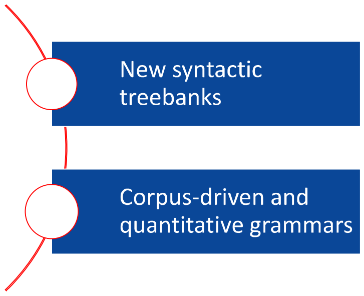
Santiago Herrera, Kim Gerdes, Bruno Guillaume, Sylvain Kahane

INTRODUCTION

AUTOGRAMM IS AN ONGOING RESEARCH PROJECT THAT AIMS TO CONTRIBUTE TO **LANGUAGE DOCUMENTATION** AND **THEORETICAL LINGUISTICS,** FROM A CROSS-LINGUISTIC AND QUANTITATIVE PERSPECTIVE.

Cross-linguistic studies **require high-quality, diverse, and comparable corpora** that are rich enough to extract precise grammatical observations to enable contrastive and typological studies.

Autogramm addresses these issues by developing simultaneously:



New syntactic treebanks

Corpus-driven and quantitative grammars

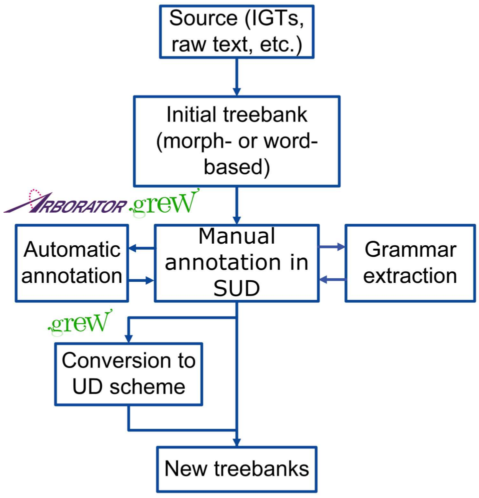
The project brings together a **heterogenous team** of field linguists, typologists, corpus annotation and formal grammar specialists, and NLP experts.

PROCESSING PIPELINE

1. Transform the source (usually interlinear glosses) into a **pre-treebank.**
2. **Syntactic annotation** can be done at the level of words or morphs
3. We use ArboratorGrew's bootstrapping system for **automatic annotation**
4. In parallel, we build **quantitative and corpus-driven grammars**

Better suited for extracting **surface rules**, especially **word order rules**.

SUD <-> UD conversion favors **treebank homogenization** and **error mining.**



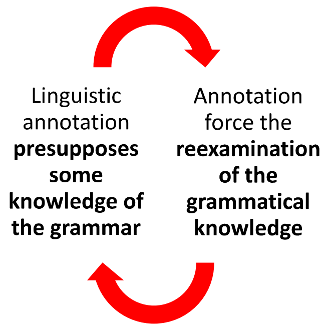
Treebanks for **Beja** and **Zaar** are already in UD database

NEW TREEBANKS

MORE THAN 15 TREEBANKS ARE UNDER DEVELOPMENT

- **Amdo Tibetan (Sinotibetan)**
- **Arabic dialects (Moroccan, Egyptian, Tunisian; Semitic)**
- **Bambara (Manding)**
- **Breton (Indo-European)**
- **Gbaya (Ubanguian)**
- **Haitian (Creole)**
- **Hausa (Chadic)**
- **Salar (Turkic)**
- **Sungwadia (Austronesian)**
- **Tuwari (Papua)**
- **Vietnamese (Austrasiatic)**
- **Yali (Papua)**
- **Ye'kwana (Carib)**
- **Etc.**

GRAMMAR EXTRACTION



Linguistic annotation **presupposes some knowledge of the grammar**

Annotation force the **reexamination of the grammatical knowledge**

Their simultaneous development could help to **reduce working time and improve the quality** of both resources

CORPUS-DRIVEN GRAMMARS

Quantitative information

- Frequency and other continuous measures

Observations ranked by relevance

- To understand the importance of each properties extracted

Variable fine-grained descriptions

- Grammars of different sizes
- Information encoded at different levels

Combinatorial explosion
Incongruent results due to annotation
Unbalanced samples

QUANTITATIVE TYPOLOGY

Compare these quantitative grammatical observations across languages and corpora

- To know exactly what makes one language different from another
- To calculate the degree to which the same observation differs between languages

Typological database

Autogramm
projet anr®

For a full list of members, future events and tools in development, visit https://autogramm.github.io/

abstract

UniDive COST Action CA21167 - Relevant groups : **WG1 and WG4**