BUILDING AN ENDANGERED LANGUAGE RESOURCE IN THE CLASSROOM

UNIVERSAL DEPENDENCIES FOR KAKATAIBO

Roberto Zariquiey (PUCP), Claudia Alvarado (PUCP), Ximena Echevarría (PUCP), Luisa Gómez (PUCP), Rosa Gonzales (PUCP), Mariana Illescas (PUCP), Sabina Oporto (PUCP), Frederic Blum (Humboldt-Universität zu Berlin and Leibniz-Zentrum Allgemeine Sprachwissenschaft), Arturo Oncevay (University of Edinburgh), Javier Vera (Pontifícia Universidade Católica de Valparaíso)

2022
1. Introduction

2. The Kakaibo Language

3. Methodology in the Classroom

4. The Kakataibo Treebank

5. Experimentation

6. Conclusions

7. Bibliographical References
2. THE KAKATAIBO LANGUAGE

1. TYPOLOGY
Mainly, an SOV and postpositional language, with a tendency to synthetic and agglutinative structures.

2. PANOAN LANGUAGE
In the experimental phase we validated this to test an automatic labeling model that took Shipibo (related) as a training language.

3. STUDIED LANGUAGE
Important: we start from the sentences documented in the grammar of the language.
3. METHODOLOGY IN THE CLASSROOM
3.1 BACKGROUND: COURSE GOALS AND STUDENTS

1. Designed for advanced undergraduate students with extensive knowledge in linguistics.

2. It had no prerequisites as it is an optional course.

3. Students with minimal technology background, and some experience in web development.
3.2 COURSE CONTENT

THEORETICAL LECTURES
Introduction to global linguistic initiatives using computational tools such as Universal Morphology or Universal Dependencies.

PRACTICALS
Guided programming exercises sing Python in Jupiter Notebook and Annotatrix, created by Tyers, Sheyanova and North for UD.

EXPERIMENTATION
Building of a new Universal Dependencies Treebank for a Peruvian minority language: Kakataibo.
3.3 COLLABORATIVE METHODOLOGY FOR THE DEVELOPMENT OF THE LANGUAGE RESOURCE

1. Lexical Segmentation
2. Root Identification
3. Part-of-Speech Identification and Tagging
4. Dependencies Generation
5. Annotation Compiling
BEFORE

(368) *Ami ka *ën piti nan!
   a=mi(‘ki)   ka *ën piti nan
that=IMPR.LOC NAR 1SG=GEN food:ABS PUT:IMPF
‘Put my food around there!’

(369)

(370) [hotel]=u(ki)  ko_a *ën xukën u-t-a
hotel=IMPR.DIR NAR:3 1SG=GEN brother:ABS come-IPFV-NON.PROX
‘My brother is coming in the direction of the hotel.’

AFTER

\[
\begin{align*}
\text{ROOT} & \rightarrow <\text{aux:sgen}> \\
 & \rightarrow <\text{nsbj:bound}> \\
 & \rightarrow <\text{obj}> \\
\text{case} \rightarrow \text{patron} \rightarrow \text{NOUN} \\
\text{case} \rightarrow \text{=batan} \rightarrow \text{PART} \\
\text{ka} \rightarrow \text{3} \rightarrow \text{PART} \\
\text{=a} \rightarrow \text{4} \rightarrow \text{PART} \\
\text{balata} \rightarrow \text{5} \rightarrow \text{NOUN} \\
\text{a-ake-x-in} \rightarrow \text{6} \rightarrow \text{VERB} \\
\text{punct} \rightarrow \text{7} \\
\end{align*}
\]
3.4 DISCUSSION

The implemented methodology can be replicated in future collaborative creation of treebanks based on grammars’ examples in the frame of NLP, programming or computational courses or workshops for Linguistics students.
4. THE KAKATAIBO TREEBANK
Framework for consistent annotation of grammar (POS, morphology features and syntactic dependencies) across world’s languages

- **Part-of-speech**: from the set of 17 POS tags, 15 of them were used to elaborate the Kakataibo treebank. **Enclitics** were labelled as **PART**.

- **UD dependencies**: 27 (of the 37 UD relations) dependencies have been used for the annotation. Subtypes of relations!
Language-specific dependencies: aux subtypes

A Kakataibo sentence featuring the dependencies **aux:sgen**, **aux:ev** and **aux**.

"It is said that, even though (he) was a blind (person), the woman did not mistreat him."
Language-specific dependencies: subtypes of the subject of a clause

A Kakataibo sentence featuring the dependencies `nsubj:free` and `nsubj:bound`. 

'\textit{I brought this fish.}'
5. EXPERIMENTATION
TASKS: POS TAGGING & DEPENDENCY PARSING

A
Monolingual training (Kakataibo only)

B
Zero-shot or fine-tune with Shipibo-Konibo
The Shipibo-Konibo (another Pano lang.) treebank did not help more, but...

Zero-shot results suggest that it could be useful for the initial annotation workflow of a new related language.

Check the F1 scores for each POS in the paper!
DEPENDENCY PARSING

KTK=Kazakh
SHP=Shipibo-Konibo

Unlabelled Attachment Score (UAS) measures the assignment of a head for any element, without considering the POS.
TRANSFER DIDN’T HELP!

Potential reasons:
- Different utterance lengths
- Different domains

But zero-shot POS tagging and delexicalized transfer for dep. parsing are promising!
6. CONCLUSIONS
FUTURE POSSIBILITIES

- Keep on improving the linguistic structure's annotation of Kakataibo
- Increase the size of our database in future lessons of the Computational Linguistics course
- Establish a replicable methodology to create treebanks for other documented languages
- Further experiment with related Panoan languages
- Continue developing other NLP tools from our treebank, such as automatic translators and predictive text
- Increase the internet presence and representativity of minority language speakers
7. BIBLIOGRAPHICAL REFERENCES


7. BIBLIOGRAPHICAL REFERENCES


7. BIBLIOGRAPHICAL REFERENCES


7. BIBLIOGRAPHICAL REFERENCES


Thank you for your attention.

Any questions?