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-Universit¨at zu Berlin and Leibniz-Zentrum Allgemeine Sprachwissenschaft), Arturo Oncevay (University of Edimburgh), Javier Vera (Pontificia Universidad Católica de Valparaíso)

2022

CONTENT

2. THE KAKATAIBO LANGUAGE

1

TYPOLOGY

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

2

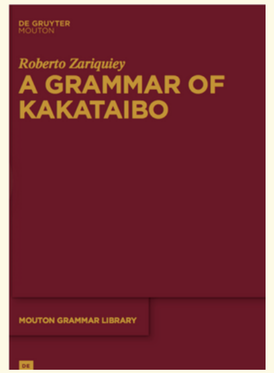
PANOAN LANGUAGE

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

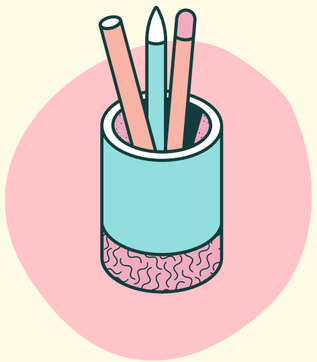
DESIGNED FOR ADVANCED UNDERGRADUATE STUDENTS WITH EXTENSIVE KNOWLEDGE IN LINGUISTICS

2

IT HAD NO PREREQUISITES AS IT IS AN OPTIONAL COURSE

3

STUDENTS WITH MINIMAL TECHNOLOGY BACKGROUND, AND SOME EXPERIENCE IN WEB DEVELOPMENT



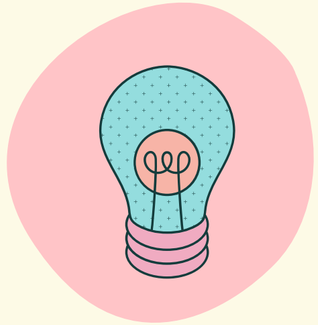
THEORETICAL LECTURES

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

3.2 COURSE CONTENT



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!*

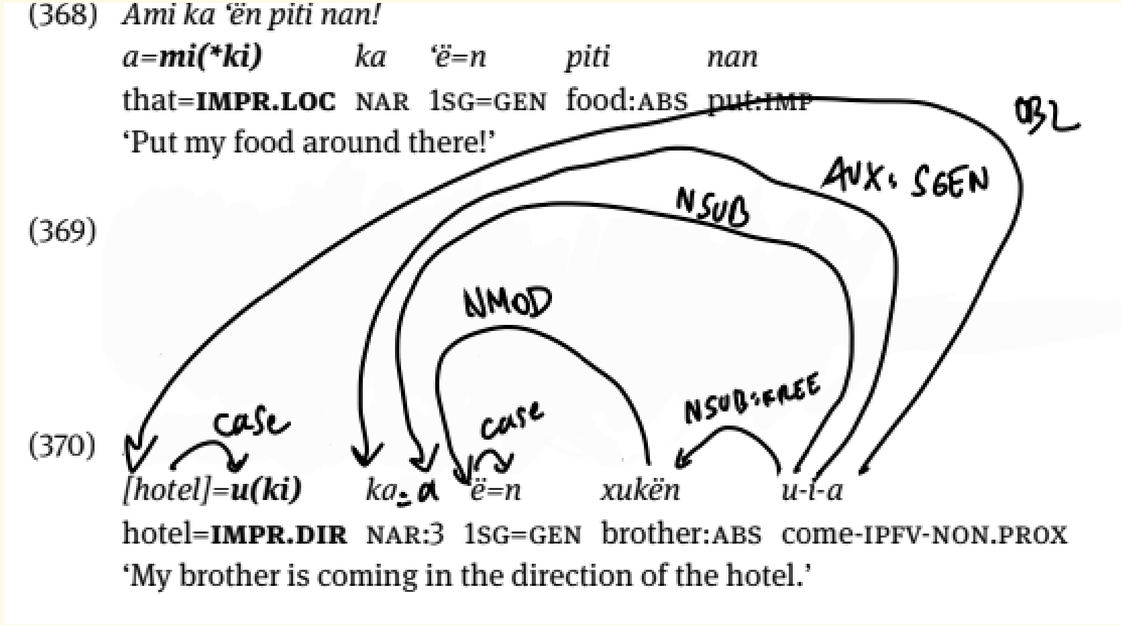
*a=**mi(*ki)*** *ka* *‘ë=n* *piti* *nan*

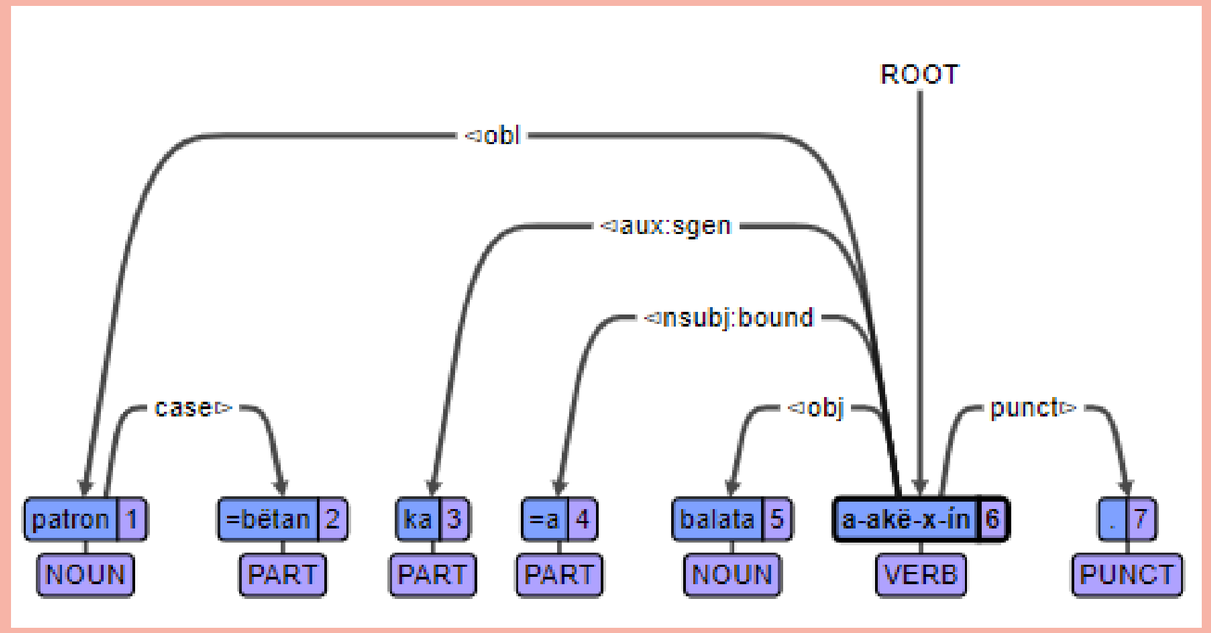
that=**IMPR.LOC** NAR 1SG=GEN food:ABS ~~put:IMP~~

‘Put my food around there!’

(369)

(370)





AFTER

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.

Inter-annotator agreement

Establishment of parameters

Use of high-quality grammars

4. THE KAKATAÏBO TREEBANK

Universal Dependencies

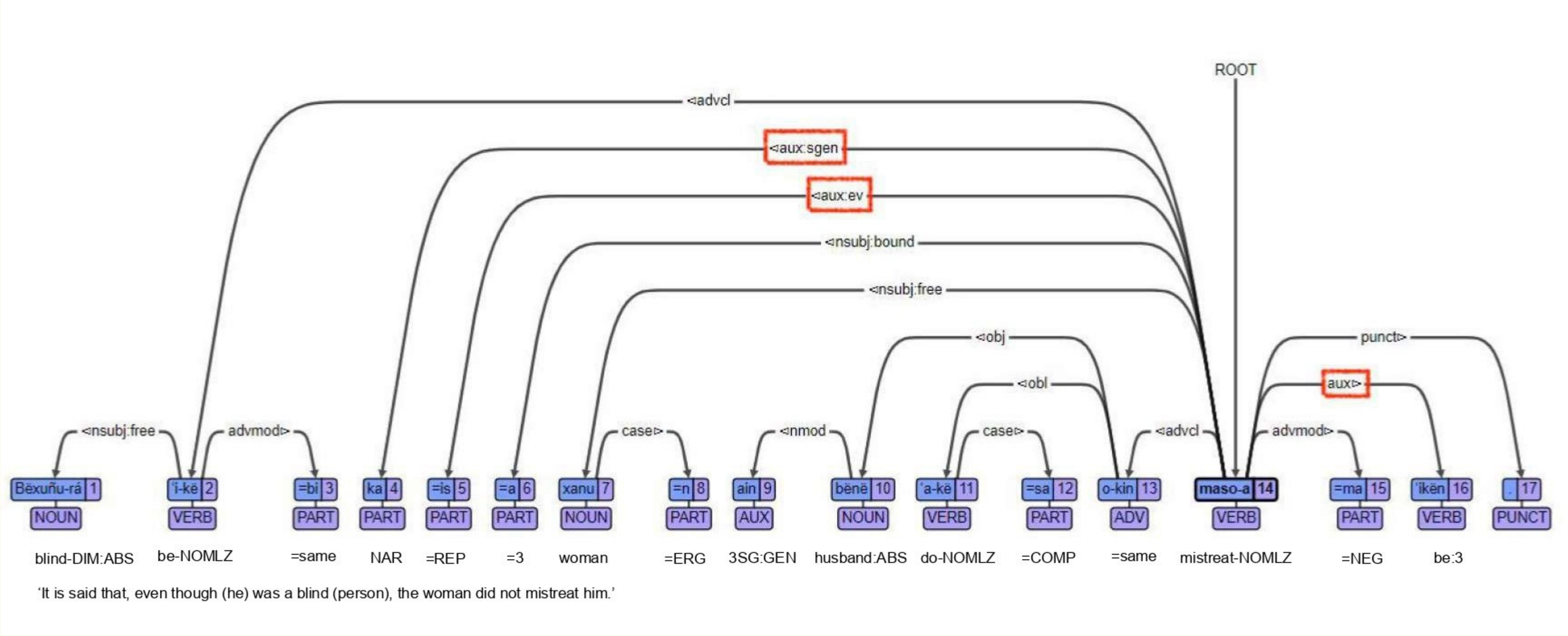
https://universaldependencies.org/guidelines.html

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**.

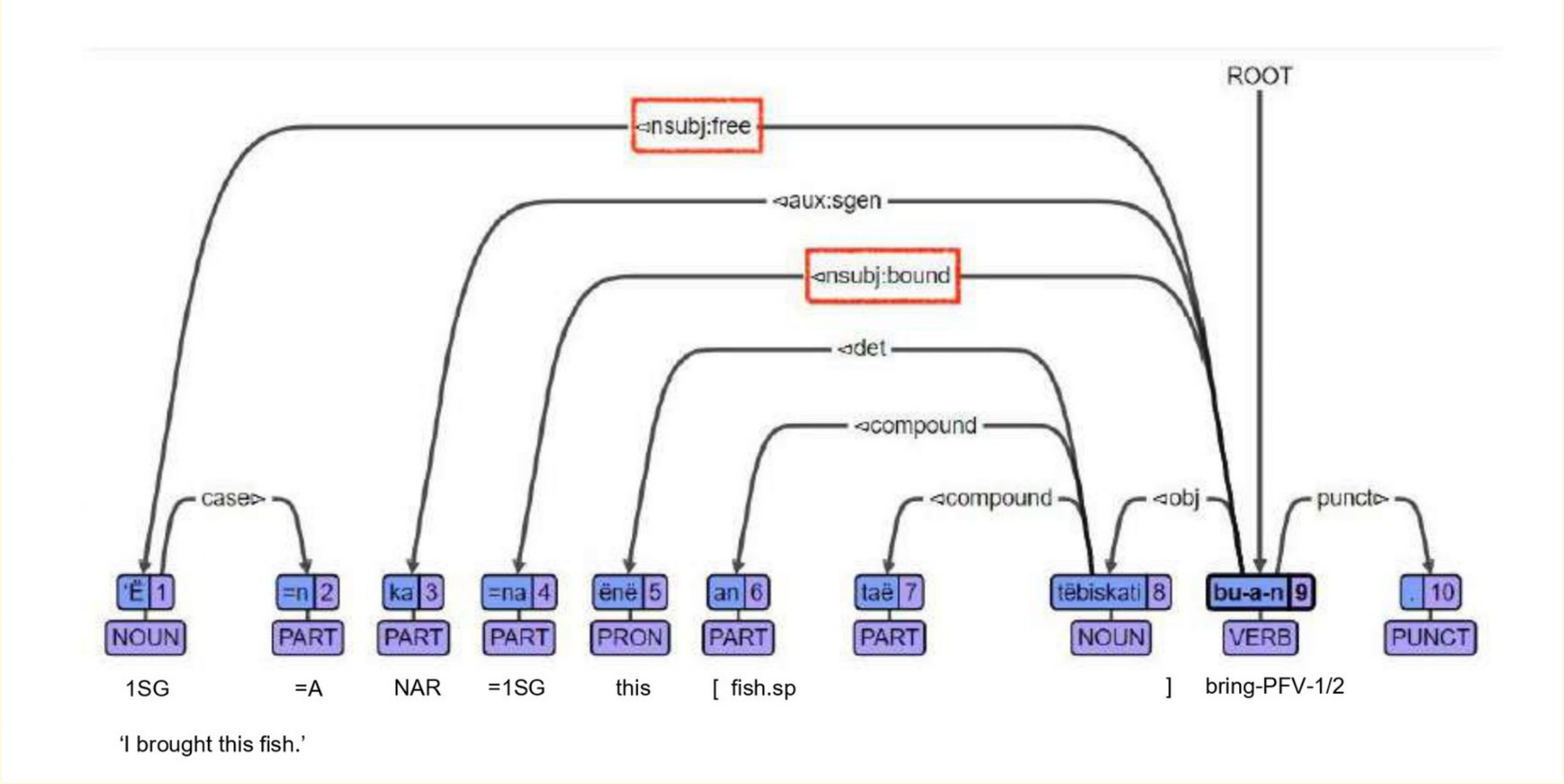
- **UD dependencies:** 27 (of the 37 UD relations) depencies have been used for the annotation. Subtypes of relations!

Language-specific dependencies: aux subtypes



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

Language-specific dependencies: subtypes of the subject of a clause



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

5. EXPERIMENTATION

TASKS:
POS TAGGING & DEPENDENCY PARSING

A

Monolingual training (Kakataibo only)

vs

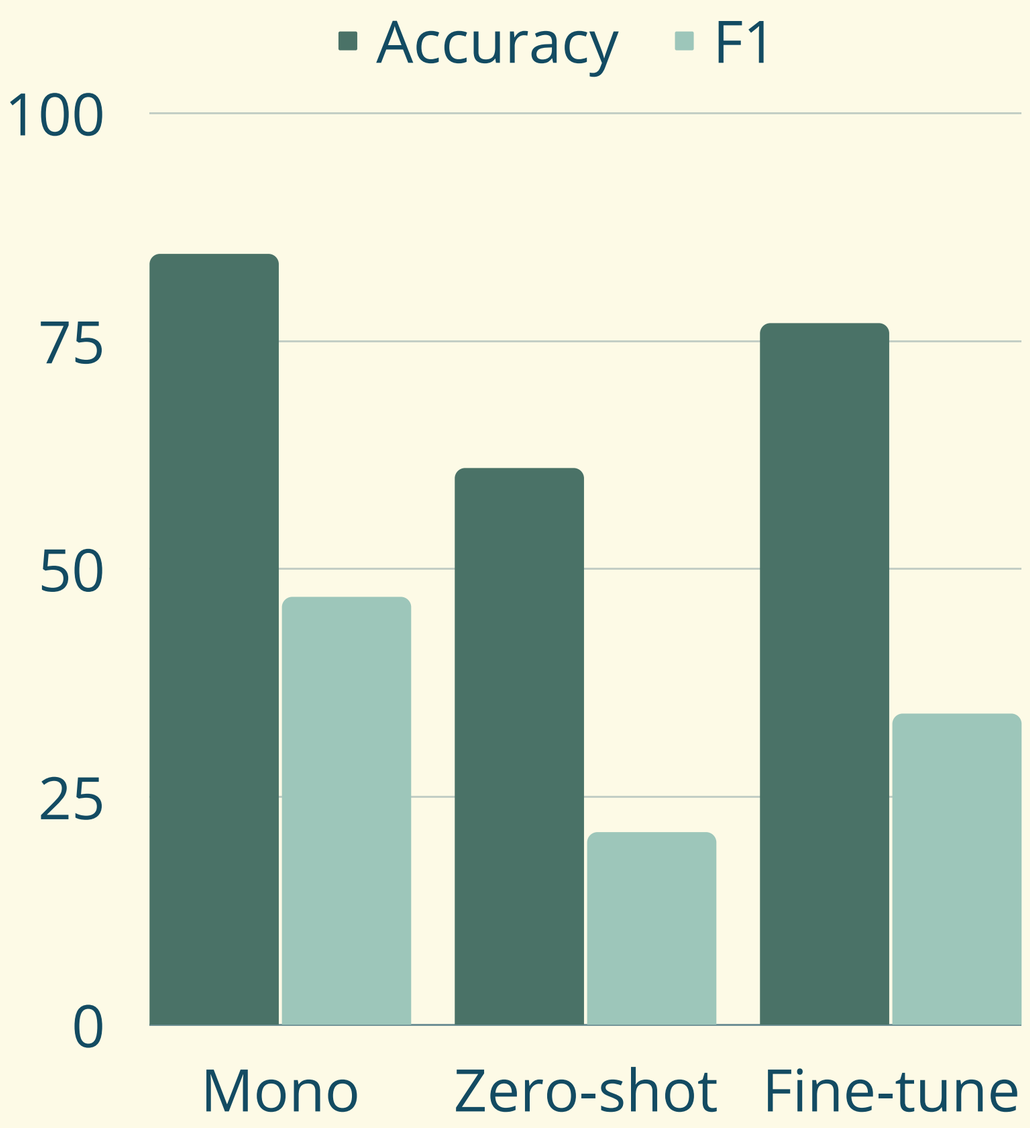
B

Zero-shot or fine-tune with Shipibo-Konibo

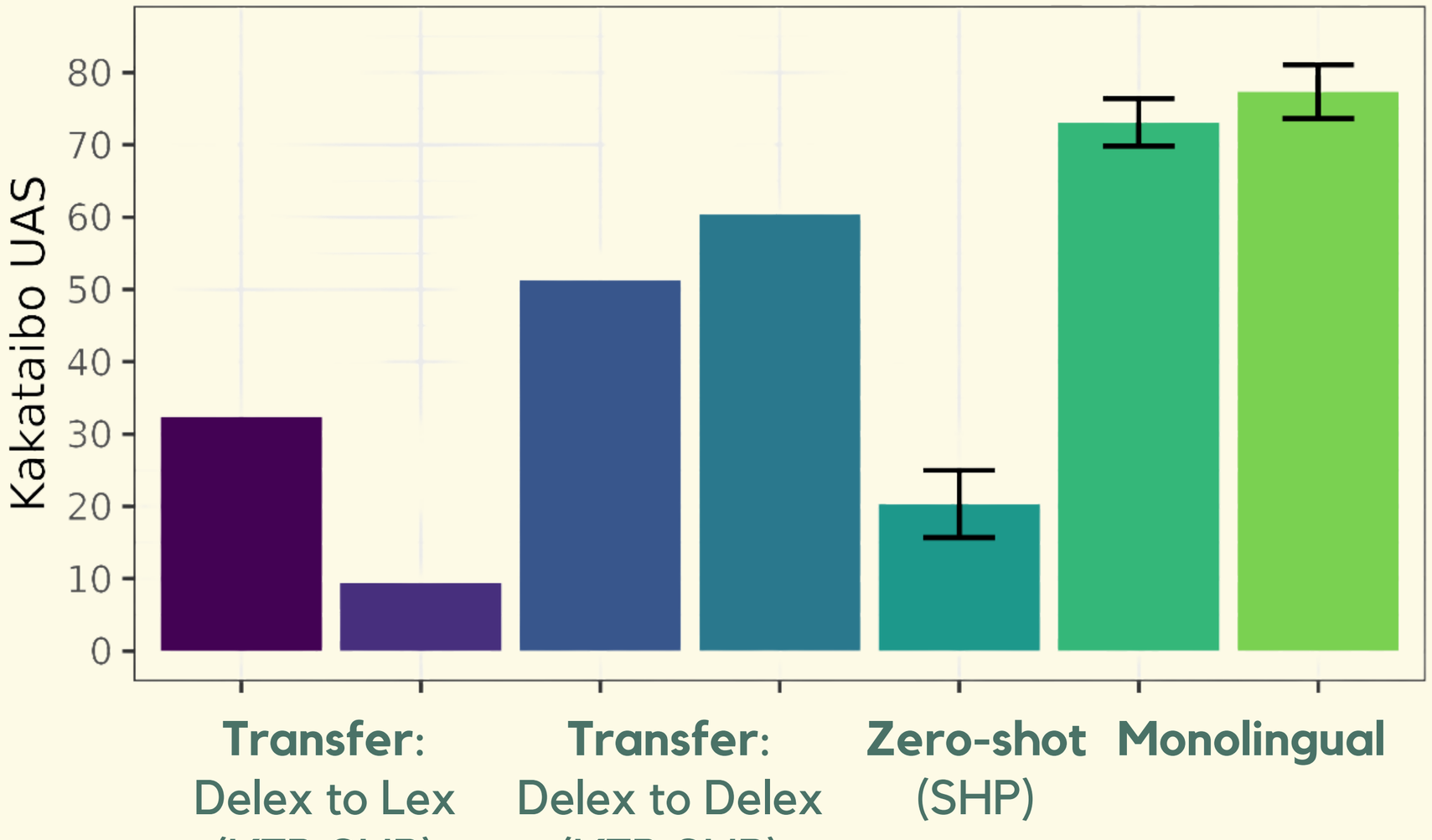
POS TAGGING

- 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



KTB=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 POSIBILITIES

- 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

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art NLP. In NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59.

Bender, E. M. (2007). Combining research and pedagogy in the development of a crosslinguistic grammar resource. In Tracy Holloway King et al., editors, Proceedings of the GEAF07 Workshop, pages 26–45, Stanford, CA. CSLI.

Bender, E. M. (2009). Linguistically naïve!=language independent: Why nlp needs linguistic typology. In Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?, pages 26–32.

7. BIBLIOGRAPHICAL REFERENCES

Croft, W., Nordquist, D., Looney, K., and Regan, M. (2017). Linguistic typology meets universal dependencies. In TLT, pages 63–75.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 4585–4592, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2):255–308, 07.

Dozat, T. and Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In ICLR 2017.

Hiippala, T. (2021). Applied language technology: NLP for the humanities. In Proceedings of the Fifth Workshop on Teaching NLP, pages 46–48, Online, June. Association for Computational Linguistics.

Hinrichs, E., Hinrichs, M., K¨ubler, S., and Trippel, T. (2019). Language technology for digital humanities: introduction to the special issue. Language Resources and Evaluation, 53(4):559–563, Dec.

7. BIBLIOGRAPHICAL REFERENCES

Makazhanov, A., Sultangazina, A., Makhambetov, O., and Yessenbayev, Z. (2015). Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report. In 3rd International Conference on Turkic Languages Processing, (TurkLang 2015), pages 338–350.

Pieter Muysken, et al., editors. (2016). South American Indigenous Language Structures (SAILS) Online. Max Planck Institute for Evolutionary Anthropology, Jena.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Pereira-Noriega, J., Mercado-Gonzales, R., Melgar, A., Sobrevilla-Cabezudo, M., and Oncevay-Marcos, A. (2017). Ship-lemmatagger: Building an nlp toolkit for a peruvian native language. In Kamil Ekˇstein et al., editors, Text, Speech, and Dialogue, pages 473–481, Cham. Springer International Publishing.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).

7. BIBLIOGRAPHICAL REFERENCES

Stenström, E. (2016). conllu. https://github.com/EmilStenstrom/conllu/.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015). A universal feature schema for rich morphological annotation and finegrained cross-lingual part-of-speech tagging. In Cerstin Mahlow et al., editors, Systems and Frameworks for Computational Morphology, pages 72–93, Cham. Springer International Publishing.

Tessmann, G. (1930). Die Indianer Nordost-Perus: grundlegende Forschungen für eine systematische Kulturkunde, volume 2 of Veröffentlichung der
Harvey-Bassler-Stiftung. Hamburg, Hamburg.

Tyers, F. M. and Washington, J. N. (2015). Towards a Free/Open-source Universal-dependency Treebank for Kazakh. In 3rd International Conference on Turkic Languages Processing, (TurkLang 2015), pages 276–289.
Tyers, F. M., Sheyanova, M., and Washington, J. N. (2018). Ud annotatrix: An annotation tool for universal dependencies. In Proceedings of the 16th International Workshop on Treebanks and LinguisticTheories (TLT16), pages 10–17.

Vajjala, S. (2021). Teaching NLP outside linguistics and computer science classrooms: Some challenges and some opportunities. In Proceedings of the FifthWorkshop on Teaching NLP, pages 149–159, Online, June. Association for Computational Linguistics.

7. BIBLIOGRAPHICAL REFERENCES

Vasquez, A., Ego Aguirre, R., Angulo, C., Miller, J., Villanueva, C., Agić, Ž., Zariquiey, R., and Oncevay, A. (2018). Toward Universal Dependencies for Shipibo-konibo. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 151–161, Brussels, Belgium, November. Association for Computational Linguistics.

Zariquiey, R., Hammarström, H., Arakaki, M., Oncevay, A., Miller, J., García, A., and Ingunza, A. (2019). Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el Perú: hacia un estado de la cuestión. Lexis, 43(2):271–337. Zariquiey, R. (2011). Aproximación dialectológica a la lengua cashibo-cacataibo (pano). Lexis, 35(1):5–46.

Zariquiey, R. (2013). Tessmann's nokamán: a linguistic investigation of a mysterious panoan group. Cadernos de Etnolingüística, 5(2):1–48.

Zariquiey, R. (2018). A grammar of Kakataibo, volumen 75 of Mouton Grammar Library. Mouton, Berlin.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In Proceedingsof the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

Zhang, Y., Li, Z., and Min, Z. (2020). Efficient Second-Order TreeCRF for Neural Dependency Parsing. In Proceedings of ACL, pages 3295–3305.

THANK YOU FOR YOUR ATTENTION

ANY QUESTIONS?