# Wolof Universal Dependency Parsing

Bill Dyer

University of Florida

April 10th 2021

# Wolof

# Wolof

- North Atlantic Niger-Congo Language
- Lingua franca in Senegal (Mc Laughin, 2008)
- Wolof dominates spoken language
- French dominates writing, so the corpora of standardized written Wolof is small
- Sentences in the Wolof UD treebank are taken from websites

# The Wolof UD Treebank

- The UD treebank annotation has 10 columns for each word:
    - number of the word in each sentence
    - word
    - lemma
    - universal part of speech
    - language-specific part of speech
    - features
    - number of the word that this word is a dependent of
    - dependency label
    - combination of dependency number and label
    - SpaceAfter word value (true or false)

# The Wolof UD Treebank

- ▶ This study depends on the Wolof UD Treebank, created by Cheikh Bamba Dione (2019)
- ▶ In this study, I edit the language specific parts of speech, dependencies, and dependency labels
- ▶ I take the edited treebanks and build a parser moder with each
- ▶ Each model has a test and train treebank, and the training data goes through the parser
- ▶ The the accuracy of the training and test data is measured for Dione's baseline and each of the edited models
- ▶ Accuracy is compared to see which parser model was most accurate

# Hypothesis

- Minimalist syntactic principles can inform Universal Dependency structures
  - Verbal morphology in Wolof does not act as a copular verb
    - Copulas are null in most copular constructions
  - Relative pronouns in headed relative clauses and free relative clauses represent the same functional head
    - Free relative clauses must contain null element

# Hypothesis

- Dependency structures inspired by minimalist syntactic analysis can improve computational parsing of a language
  - Parser models trained on treebanks influenced by minimalist syntactic analysis will show improved accuracy from the baseline model

# Determiners

(1)    a.    *nit*    ***k-i***
           person    KClass-the

           'the person'

     b.    *cin*    ***l-i***
           large.pot    LClass-the

           'the large pot'

     c.    *cin*    ***y-a***
           large.pot    YClass-those

           'those large pots'

     d.    *jamono*    ***j-ooju***
           era        JClass-that.far

           'that time long ago'

# Headed Relative Clauses

(2) a. *téere **b**-**u***     *am solo*
    book BClass-a has importance

    'a book that is important'

   b. *làkk*     ***y**-**ii***     *ñu nàmp*
      language YClass-these they nurse

      'these languages here that they are raised speaking'

   c. *jiggéen-i Afrik*   ***y**-**ooyu***     *dekkal*
      women-of Africa YClass-those.far bring.back

      *cosaan-i maam*
      tradition-of grandparent

      'those women of Africa there who bring back the traditions of their grandparents'

# Free Relative Clauses

(3) a. ***k-i***      *leen*    *taxawal-oon*
ClassK-the them stand.up.against-PAST

    'the one who stood up against them'

   b. ***ñ-a***      *ànd-oon*      *ak*    *Ablaay*
ClassÑ-those accompany-PAST with Abdoulaye
    *Wàdd*
    Wade

    'those who accompanied Abdoulaye Wade'

# Copular Constructions

(4) a. Kolle sama mag la.
   *Kolle sama mag       la*
   Kolle my   older.sibling OBJ.FOCUS

   'Kolle is my OLDER SISTER"

   b. Kolle mooy sama mag.
   *Kolle mu-a-di           sama mag*
   Kolle she-SUBJ.FOCUS-is my   older.sibling

   'KOLLE is my older sister'

# Imperfect *di*

(5)   a.   *Kànj la-a    jënd.*
           okra FOC-I  sell

           'I have sold OKRA.'

      b.   *Kànj la-a-y    jënd.*
           okra FOC-I-am  sell

           'I'm selling OKRA.'

      c.   *Abdu  mu-a    jënd kànj.*
           Abdou  he-FOC  sell  okra

           'ABDOU has sold okra.'

      d.   *Abdu mu-a-y    jënd kànj.*
           Abdu  he-FOC-is  sell  okra

           'Abdu is selling okra.'
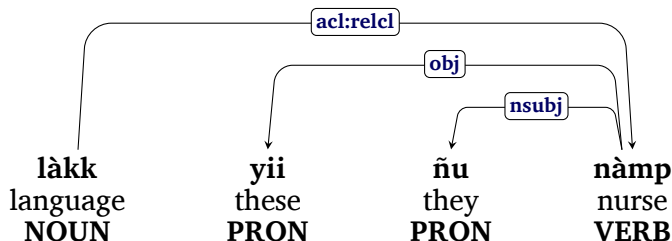
# Headed Relative Clause Dependencies

(6) làkk yii ñu nàmp
'these languages here that they are raised speaking'



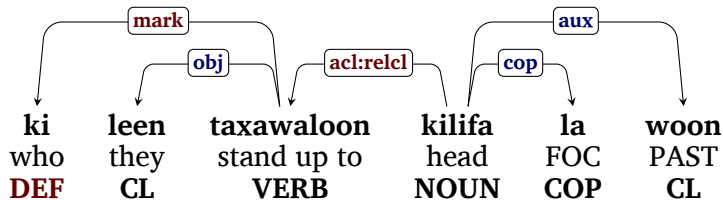| làkk | yii | ñu | nàmp |
|------|------|------|------|
| language | these | they | nurse |
| **NOUN** | **PRON** | **PRON** | **VERB** |

# Free Relative Clause Dependency

(7) ki leen taxawaloon kilifa la woon
  'the one they stood up against was a head of household"
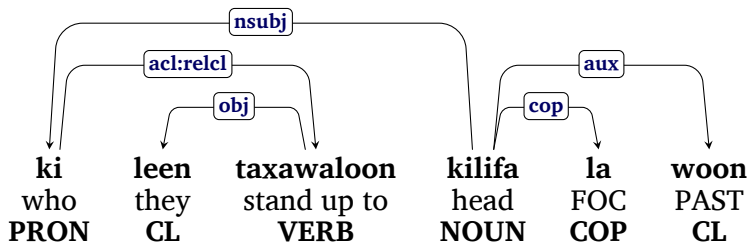
# Free Relative Clauses Reassigned with *mark* label

(8) ki leen taxawaloon kilifa la woon
   'the one they stood up against was a head of household"

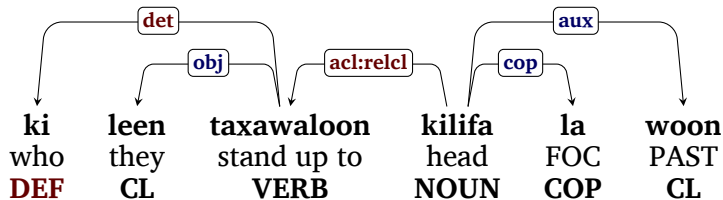# Headed Relative Clauses Reassigned with *mark* label

(9) làkk yii ñu nàmp
'these languages here that they are raised speaking'



| **làkk** | **yii** | **ñu** | **nàmp** |
| language | these | they | nurse |
| **NOUN** | **DEF** | **PRON** | **VERB** |
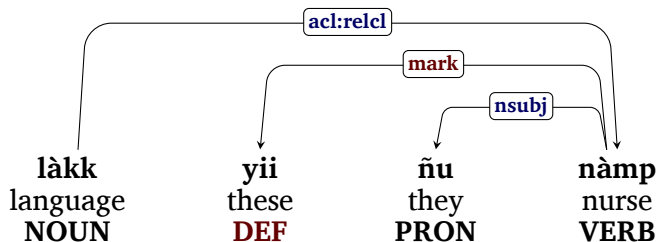
# Free Relative Clauses Reassigned with *det* label

(10)  ki leen taxawaloon
      'who they stood up against'



| **ki** | **leen** | **taxawaloon** | **kilifa** | **la** | **woon** |
|--------|----------|----------------|------------|--------|----------|
| who    | they     | stand up to    | head       | FOC    | PAST     |
| **DEF** | **CL**  | **VERB**       | **NOUN**   | **COP** | **CL**  |

# Headed Relative Clauses Reassigned with *det* label

(11) làkk yii ñu nàmp
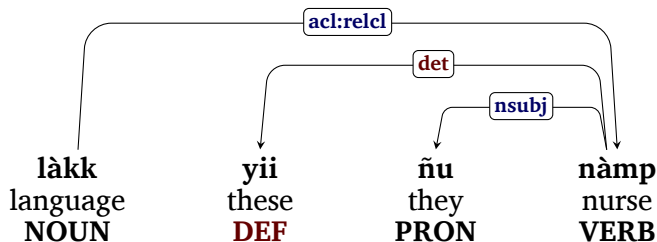      'these languages here that they are raised speaking'



| **làkk** | **yii** | **ñu** | **nàmp** |
|----------|---------|--------|----------|
| language | these   | they   | nurse    |
| **NOUN** | **DEF** | **PRON** | **VERB** |

# Copulas

Table: Tag assignment for select lemmas when not assigned COP

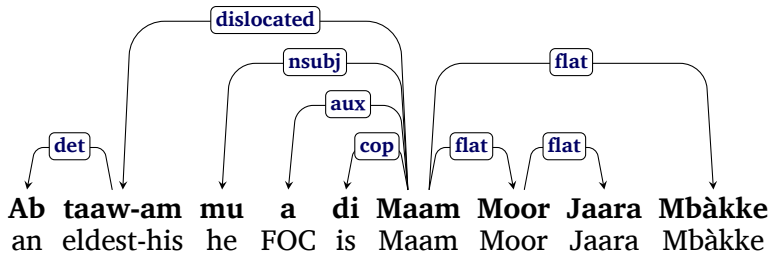| INFL | | AUX | |
|------|----------|-------|----------|
| Lemma | Function | Lemma | Function |
| la | Complement Focus | ngi | Progressive Aspect |
| da | Verbal Focus | du | Negative |
| daan | Past Habitual Aspect, focus clauses | daan | Past Habitual Aspect, non-focus clauses |
| | | di | Imperfect Aspect |

# New Labels for Lemmas Previously Labeled Copulas

Table: Category reassigned to selected lemmas previously assigned COP

| INFL | | AUX | |
|---|---|---|---|
| Lemma | Function | Lemma | Function |
| la | Complement Focus | ngi | Progressive Aspect |
| da | Verbal Focus | daan | Past Habitual Aspect |
| du | Negative | di | Imperfect Aspect |

# *Di* copula with nominal complement

(12)  Ab taawam mooy Maam Moor Jaara Mbàkke
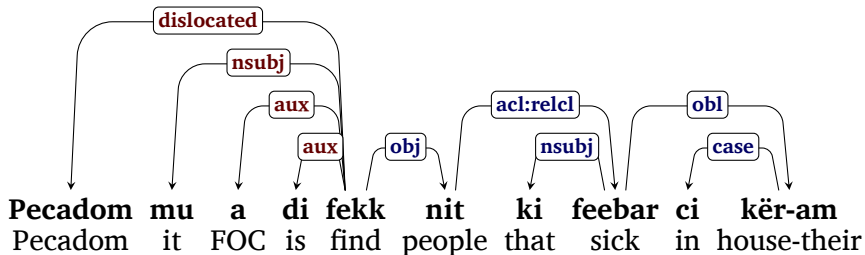      'HIS ELDEST CHILD is Maam Moor Jaara Mbàkke'



|  | **det** | | **dislocated** | **nsubj** | **aux** | **cop** | | **flat** | **flat** | **flat** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ab** | **taaw-am** | **mu** | **a** | **di** | **Maam** | **Moor** | **Jaara** | **Mbàkke** | | |
| an | eldest-his | he | FOC | is | Maam | Moor | Jaara | Mbàkke | | |

# *Di* copula with clausal complement

(13) Pecadom mooy fekk nit ki feebar ci këram.
'PECADOM finds people that are sick in their home.'



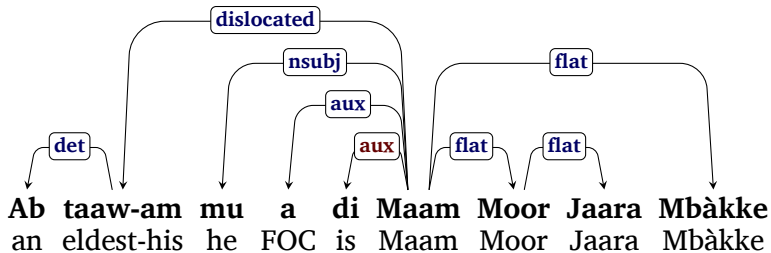| Pecadom | mu | a | di | fekk | nit | ki | feebar | ci | kër-am |
|---------|-----|-----|-----|------|--------|------|--------|-----|------------|
| Pecadom | it | FOC | is | find | people | that | sick | in | house-their |
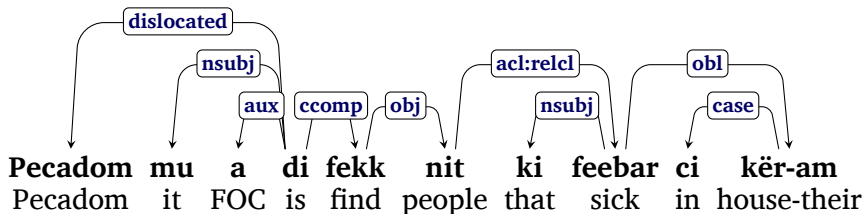
# Reassigned *di* copula with clausal complement

(14) Pecadom mooy fekk nit ki feebar ci këram.
'PECADOM finds people that are sick in their home.'

# Reassigned *di* copula with nominal complement

(15)  Ab taawam mooy Maam Moor Jaara Mbàkke
      'HIS ELDEST CHILD is Maam Moor Jaara Mbàkke'



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Ab** | **taaw-am** | **mu** | **a** | **di** | **Maam** | **Moor** | **Jaara** | **Mbàkke** |
| an | eldest-his | he | FOC | is | Maam | Moor | Jaara | Mbàkke |

# SpaCy Parser Pipeline

- ▶ .conllu UD treebanks for all three models are converted to .json format
- ▶ Separate treebanks for train, development, and test
- ▶ Four treebank sets result in four parser models:
  - ▶ Baseline is the unedited treebank (Dione, 2019)
  - ▶ All definites given DEF part of speech tag, relative pronouns treated as complementizers
  - ▶ All definites given DEF part of speech tag, relative pronouns treated as determiners
  - ▶ COP tag replaced by alternate possible tag for each lemma

# The MaltParser Parser Pipeline

- ▶ Consisted of two seperate tools: TreeTagger (Schmid, 1994) and MaltParser (Nilsson, 2008)
- ▶ The .conllu files only contained the word number, word form, lemma, universal POS tag, and Wolof POS tag for each word
  - ▶ all that could be produced by the tagger
- ▶ After a .conllu file was prepared for each model, it was used as input into MaltParser

# Accuracy for UD labels and relations with corresponding parser pipeline

| Num. | Pipeline | spaCy UD Label | spaCy Univ. Dep. | Malt UD Label | Malt Univ. Dep. |
|------|----------|----------------|------------------|---------------|-----------------|
| 0 | Baseline | 76.4% | 71.1% | 72.7% | 70.4% |
| 1 | DEF tag, RC pron as *det* | 77.9% | 71.7% | 74.9% | 72.9% |
| 2 | DEF tag, RC pron as *mark* | 77.8% | 71.4% | 74.0% | 73.2% |
| 3 | Copulas Relabeled | 77.4% | 71.2% | 73.9% | 70.7% |
| 4 | Num. 1 + Num. 3 | 78.0% | 71.4% | 76.1% | 73.3% |

# Analysis

- Model that treats the relative pronoun as an extracted determiner, not complementizer, results in more accurate parsing
- Treating definites as one part of speech category improves parsing
- The copular analysis for verbal morphology does not seem to 'fit' Wolof
- Accuracy increased for both parser pipelines, so parser alone is not responsible for improvement

# Conclusions

- Adopting a unified syntactic approach to verbal morphology and relative clauses improves accuracy in Wolof

- These improvements can inform future parsers of African langauges

# Future Research

- Build on Wolof Universal Dependency treebank with new data according to new labeling
- Creating Universal Dependency treebanks for languages with similar determiner/relative clause pronoun systems

# References

- Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

- Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.lp.hscl.ufl.edu.

- Dione, Cheikh Bamba (2019). Developing Universal Dependencies for Wolof. Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019). Paris, France

- Ka, Omar (1994). Wolof Phonology and Morphology. University Press of America. Lanham, Maryland.

# References

- Ka, Omar (1994). Nanu Dégg Wolof. National African Languages Resource Center. Madison, Wisconsin.
- Mc Laughlin, Fiona (2008). Senegal: The Emergence of a National Lingua Franca. In Simpson, Andrew (ed.), Language and National Identity In Africa, 79-97. Oxford: Oxford University Press.
- Nilsson, Jens and Joakim Nivre (2008). MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. Sixth international conference on Language Resources and Evaluation, Marrakech, Morocco.
- Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
- Torrence, Harold (2013). The Clause Structure of Wolof. John Benjamins, Philadelphia.