# Morpho-Syntactic Analyso-Parsing

Omer Goldman, Leonie Weissweiler, Reut Tsarfaty

# Preface


UNAZIGMIT
(UNAZIĠMIT)
SIRENIK
CENTRAL
SIBERIAN
YUPIK
NAUKAN
YUPIK
ST. LAWRENCE
ISLAND YUPIK
NORTON
SOUND
HOOPER BAY-
CHEVAK
GENERAL
CENTRAL
YUPʻIK
EGEGIK
KONIAG
CHUGACH
ALUTIIQ

# Preface

- What NLP tools are available for Yupik?

# Preface

- What NLP tools are available for Yupik?
- No language model, obviously

# Preface

- What NLP tools are available for Yupik?
- No language model, obviously
- But we should be able to do

# Preface

- What NLP tools are available for Yupik?
- No language model, obviously
- But we should be able to do
  - Dependency parsing

# Preface

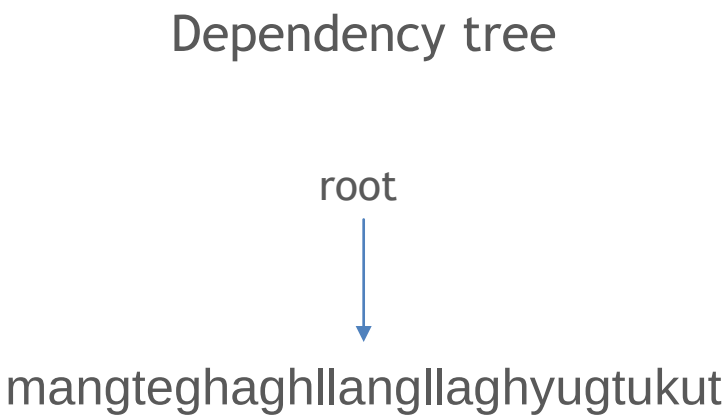
- What NLP tools are available for Yupik?
- No language model, obviously
- But we should be able to do
  - Dependency parsing
  - morphological analysis

# Preface

mangteghaghllangllaghyugtukut (*we want to make a big house)*

# Preface

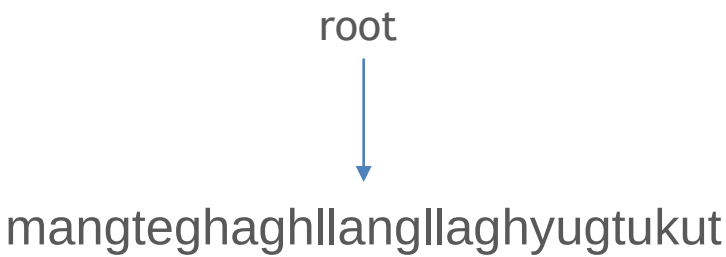
mangteghaghllangllaghyugtukut (*we want to make a big house)*



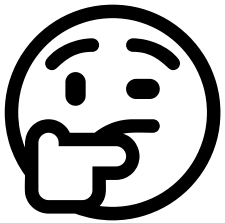
# Preface

mangteghaghllangllaghyugtukut (*we want to make a big house)*

Dependency tree




Morphological data



# Preface

- Shared tasks = resource creation

# Preface

- Shared tasks = resource creation
- Pushing the boundaries of multilingual shared tasks

→ more language diversity

# Preface

- Shared tasks = resource creation
- Pushing the boundaries of multilingual shared tasks
  → more language diversity
- Can we define a task that will be natural to these kind of languages?

# Preface

- Shared tasks = resource creation
- Pushing the boundaries of multilingual shared tasks
  → more language diversity
- Can we define a task that will be natural to these kind of languages?
- And lose nothing in the process?

# Preface

- Shared tasks = resource creation
- Pushing the boundaries of multilingual shared tasks
  → more language diversity
- Can we define a task that will be natural to these kind of languages?
- And lose nothing in the process?

Yes we can!

# The Problem

# The Problem

- A word is a sneaky little concept

# The Problem

- A word is a sneaky little concept
- “Multi-word words”

# The Problem

- A word is a sneaky little concept
- “Multi-word words”
  - Yupik: mangteghaghllangllaghyugtukut

# The Problem

- A word is a sneaky little concept
- “Multi-word words”
  - Yupik: mangteghaghllangllaghyugtukut
  - Hungarian: szeretlek (“I love you”)

# The Problem

- A word is a sneaky little concept
- “Multi-word words”
  - Yupik: mangteghaghllangllaghyugtukut
  - Hungarian: szeretlek (“I love you”)
  - Swahili: nitakichopenda (“the thing that I will love”)

# The Problem

- A word is a sneaky little concept
- “Multi-word words”
  - Yupik: mangteghaghllangllaghyugtukut
  - Hungarian: szeretlek (“I love you”)
  - Swahili: nitakichopenda (“the thing that I will love”)
- UD’s approach - segmentation

# The Problem

- A word is a sneaky little concept
- “Multi-word words”
  - Yupik: mangteghaghllangllaghyugtukut
  - Hungarian: szeretlek (“I love you”)
  - Swahili: nitakichopenda (“the thing that I will love”)
- UD’s approach - segmentation
  - Inconsistent segmentation

# Segmentation

# Segmentation

- Possible approaches:

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result
    - szeretlek → én szeretek téged

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result
    - szeretlek → én szeretek téged
  - Rephrasing → not always possible

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result
    - szeretlek → én szeretek téged
  - Rephrasing → not always possible
    - nitakichopenda → ambacho nitakipenda

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result
    - szeretlek → én szeretek téged
  - Rephrasing → not always possible
    - nitakichopenda → ambacho nitakipenda
  - More?

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result
    - szeretlek → én szeretek téged
  - Rephrasing → not always possible
    - nitakichopenda → ambacho nitakipenda
  - More?

Every treebank chooses its strategy

# Segmentation

- Possible approaches:
  - Vanilla segmentation → works only with peripheral elements
  - “Imagining” missing words → ungrammatical result
    - szeretlek → én szeretek téged
  - Rephrasing → not always possible
    - nitakichopenda → ambacho nitakipenda
  - More?

Every treebank chooses its strategy

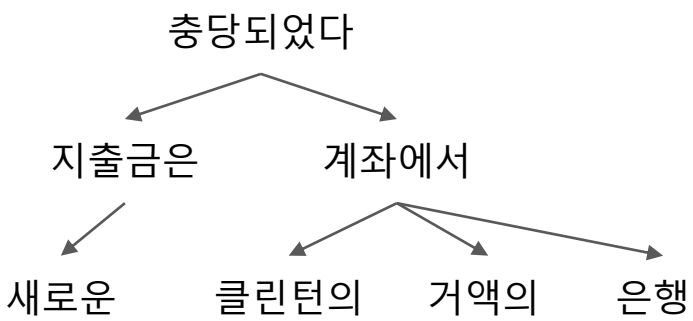
→ Inconsistent, sometimes even within a language

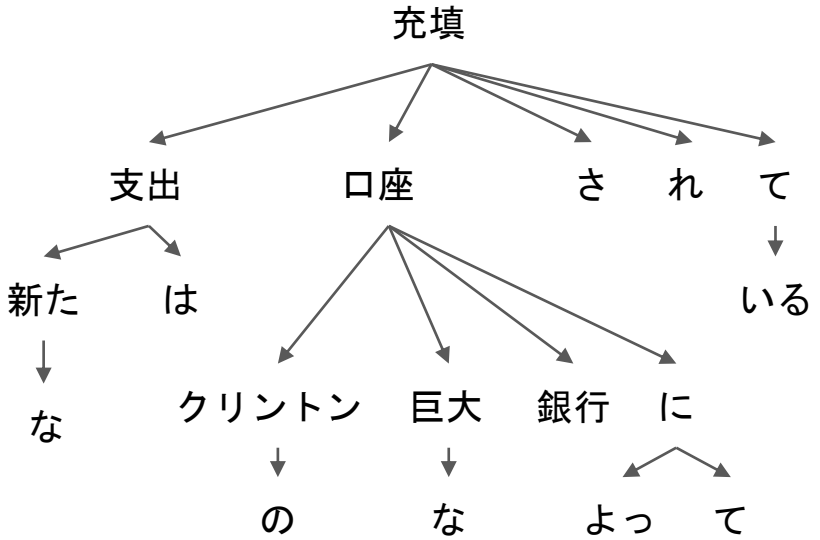
# Segmentation

- English: The new spending is fueled by Clinton’s large bank account.

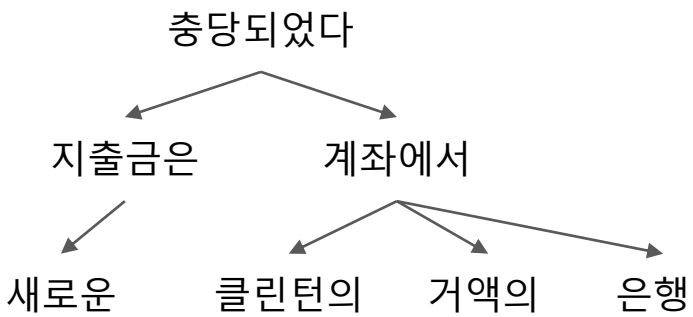
# Segmentation

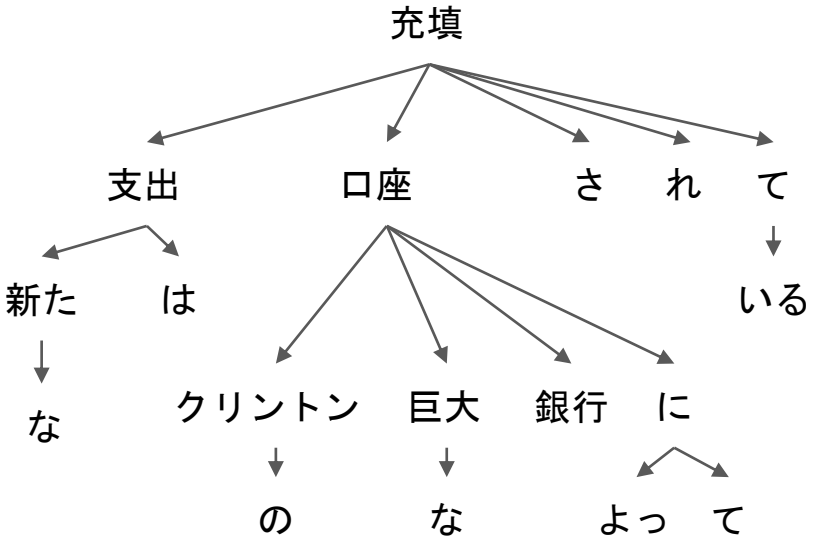
- English: The new spending is fueled by Clinton’s large bank account.



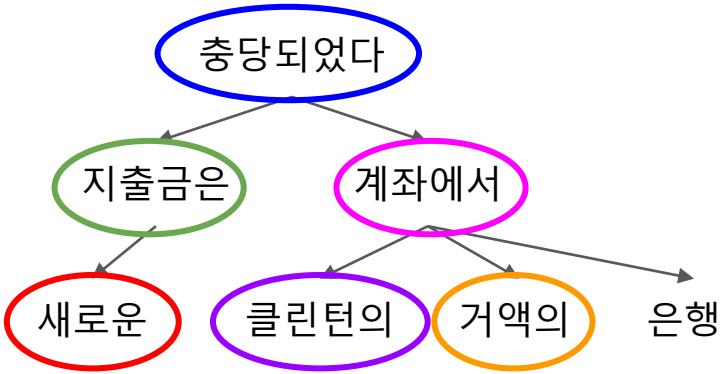



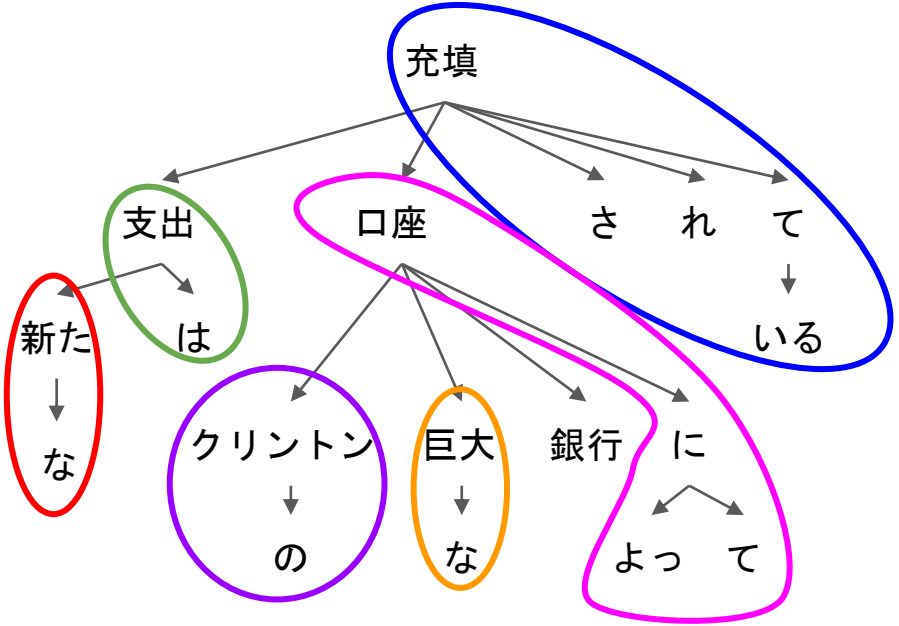
# Segmentation



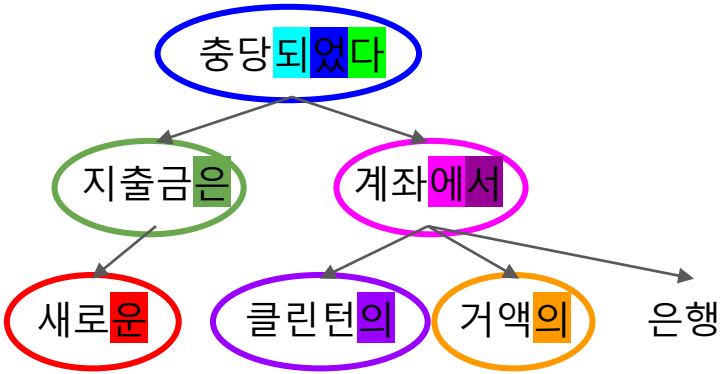



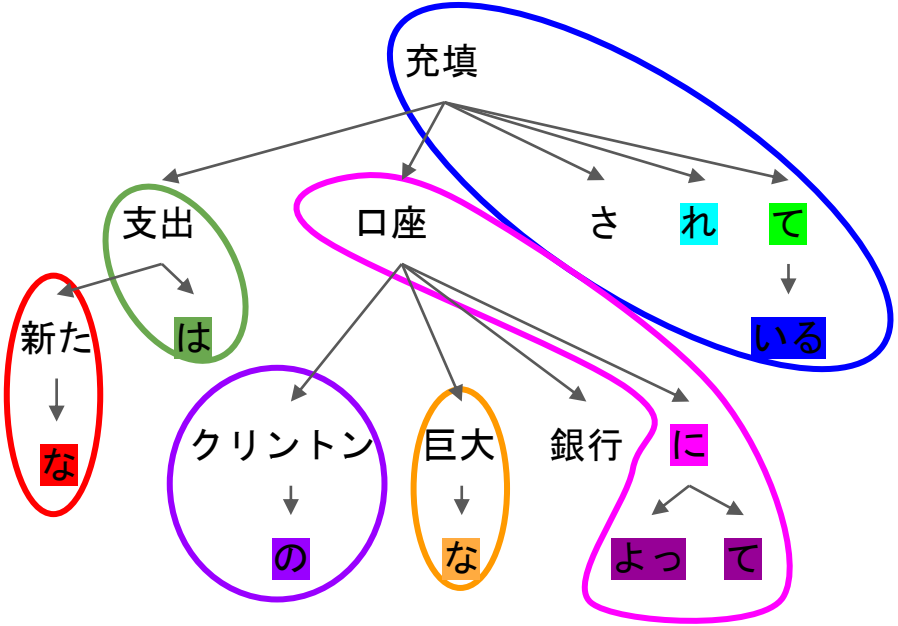
# Segmentation



# Segmentation



# Segmentation

- And what happens with Yupik?

# Segmentation

- And what happens with Yupik?
- Full morphemic segmentation

# Segmentation

- And what happens with Yupik?
- Full morphemic segmentation
  - Mangtegha-ghlla-ngllagh-yug-tu-kut

# Segmentation

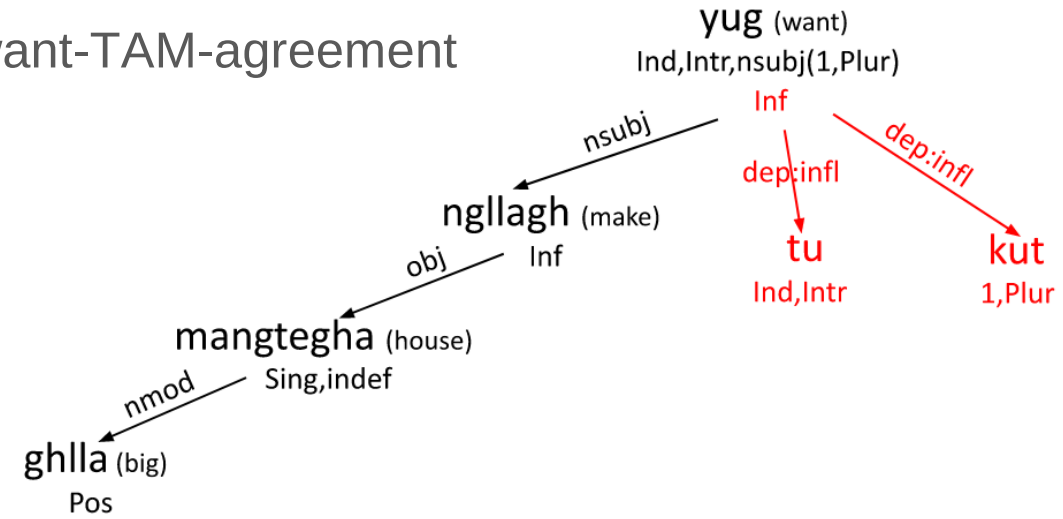
- And what happens with Yupik?
- Full morphemic segmentation
  - Mangtegha-ghlla-ngllagh-yug-tu-kut
  - House-big-make-want-TAM-agreement

# Segmentation

- And what happens with Yupik?
- Full morphemic segmentation
  - Mangtegha-ghlla-ngllagh-yug-tu-kut
  - House-big-make-want-TAM-agreement
- Is it syntax?

# Segmentation

- And what happens with Yupik?
- Full morphemic segmentation
  - Mangtegha-ghlla-ngllagh-yug-tu-kut
  - House-big-make-want-TAM-agreement
- Is it syntax?



# The Solution

# The Solution

- Ditch words!

# The Solution

- Ditch words!
- Adopt the function-content frontier as the divide between morphology and syntax

# The Solution

- Ditch words!
- Adopt the function-content frontier as the divide between morphology and syntax
  - Functions = features

# The Solution

- Ditch words!
- Adopt the function-content frontier as the divide between morphology and syntax
  - Functions = features
  - Lexemes = dependency nodes

# The Solution

- Ditch words!
- Adopt the function-content frontier as the divide between morphology and syntax
  - Functions = features
  - Lexemes = dependency nodes

- Fuse morphology and syntax into one task:

# The Solution

- Ditch words!
- Adopt the function-content frontier as the divide between morphology and syntax
  - Functions = features
  - Lexemes = dependency nodes

- Fuse morphology and syntax into one task:

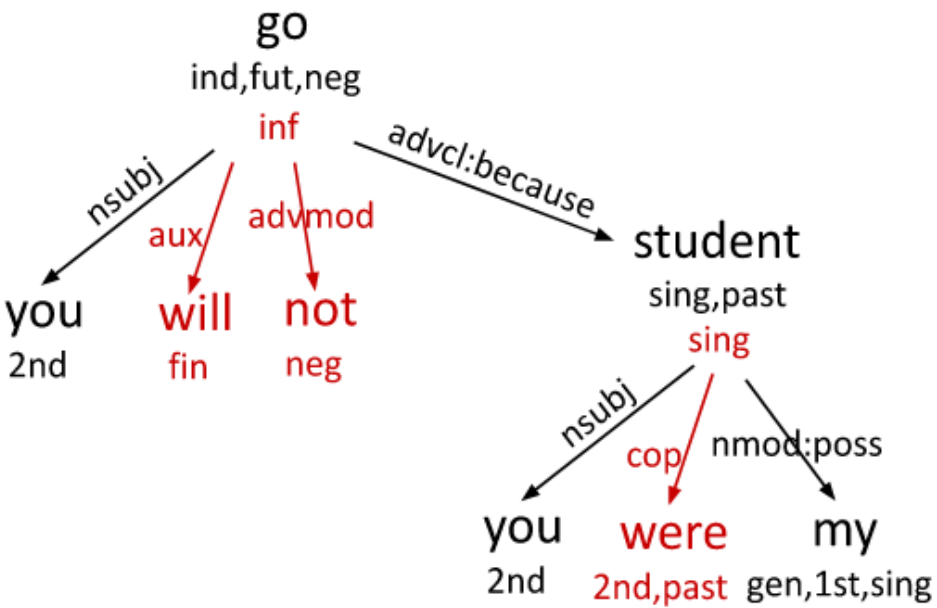
Morpho-syntactic analyso-parsing

# Morpho-Syntactic Analyso-Parsing

- English: You will not go because you were my student.
- Turkish: sen gelmeyeceksin çünkü sen benim öğrencimdin

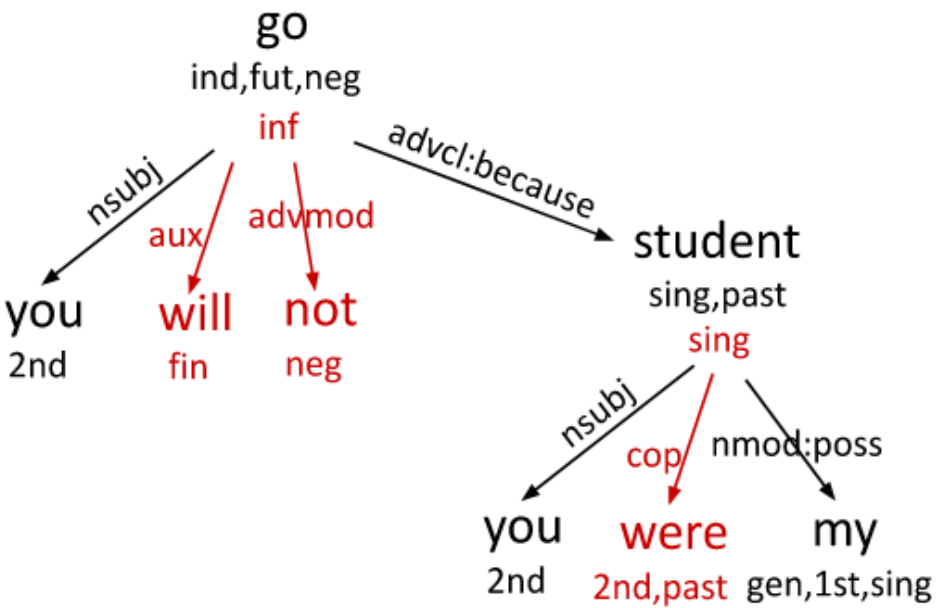
# Morpho-Syntactic Analyso-Parsing

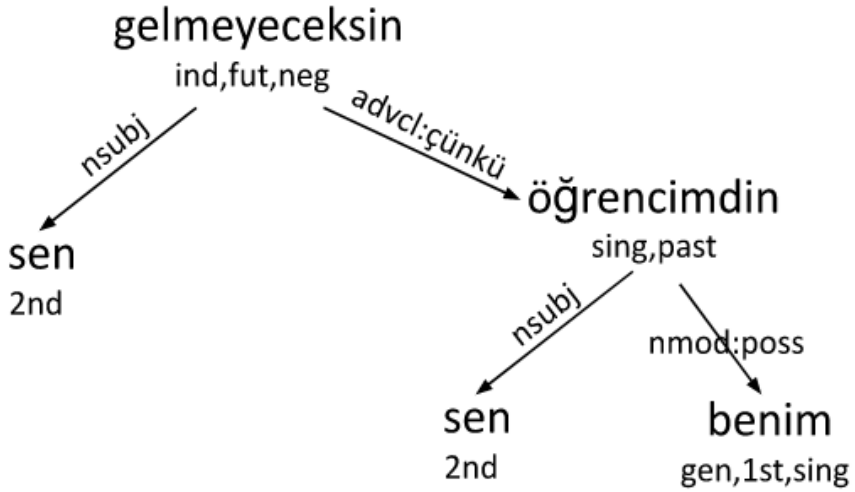
- English: You will not go because you were my student.
- Turkish: sen gelmeyeceksin çünkü sen benim öğrencimdin



# Morpho-Syntactic Analyso-Parsing

- English: You will not go because you were my student.
- Turkish: sen gelmeyeceksin çünkü sen benim öğrencimdin



# The Effects of the Suggestion

# The Effects of the Suggestion

- Surface periphrastic features for isolating languages

# The Effects of the Suggestion

- Surface periphrastic features for isolating languages
  - Every function is a feature

# The Effects of the Suggestion

- Surface periphrastic features for isolating languages
  - Every function is a feature
  - No need to decide if it’s a morpheme or an auxiliary

# The Effects of the Suggestion

- Surface periphrastic features for isolating languages
  - Every function is a feature
  - No need to decide if it’s a morpheme or an auxiliary

- Surface predicate-argument structure for polysynthetic languages

# The Effects of the Suggestion

- Surface periphrastic features for isolating languages
  - Every function is a feature
  - No need to decide if it’s a morpheme or an auxiliary

- Surface predicate-argument structure for polysynthetic languages
  - Every content lexeme is a node

# The Effects of the Suggestion

- Surface periphrastic features for isolating languages
  - Every function is a feature
  - No need to decide if it’s a morpheme or an auxiliary

- Surface predicate-argument structure for polysynthetic languages
  - Every content lexeme is a node
  - No need to separate function morphemes

# Practical Changes to CoNLL-U files

# Practical Changes to CoNLL-U files

- Not changing any existing data

# Practical Changes to CoNLL-U files

- Not changing any existing data
- Adding phrase-level features to content nodes

# Practical Changes to CoNLL-U files

- Not changing any existing data
- Adding phrase-level features to content nodes

- All nodes = syntactic tree

# Practical Changes to CoNLL-U files

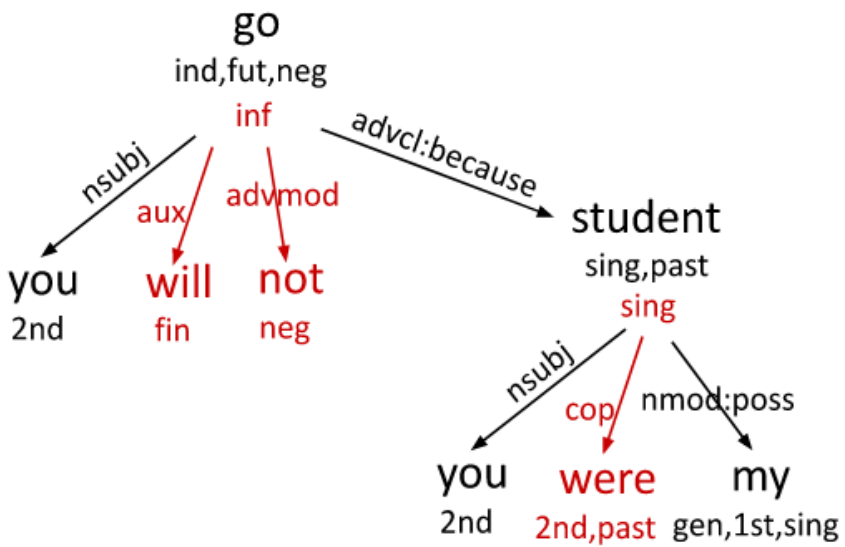
- Not changing any existing data
- Adding phrase-level features to content nodes

- All nodes = syntactic tree
- Only nodes with p-feats
  = morpho-syntactic tree

# Practical Changes to CoNLL-U files

- Not changing any existing data
- Adding phrase-level features to content nodes

- All nodes = syntactic tree
- Only nodes with p-feats
  = morpho-syntactic tree



# English Example

| ID | Form | Lemma | POS | FEATS | HEAD | DEP | P-FEATS |
|---|---|---|---|---|---|---|---|
| 1 | you | you | PRON | Nom;2;Sg | 4 | nsubj | Nom;2;Sg |
| 2 | will | will | AUX | Fin | 4 | aux | |
| 3 | not | not | PART | Neg | 4 | advmod | |
| 4 | go | go | VERB | Inf | 0 | root | Fin;Ind;Fut;Neg |
| 5 | because | because | SCONJ | - | 9 | mark | Nom;2;Sg |
| 6 | you | you | PRON | Nom;2;Sg | 9 | nsubj | |
| 7 | were | be | AUX | Fin;Ind;Past;2;Sg | 9 | cop | |
| 8 | my | my | PRON | Gen;1;Sg | 9 | nmod:poss | Gen;1;Sg |
| 9 | student | student | NOUN | Sg | 4 | advcl:because | Sg;Ind;Past |

# Modelling Morpho-Syntax

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word
- Successful model will be able to

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word
- Successful model will be able to
  - Predict predicate-argument structure in polysynthetic languages

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word
- Successful model will be able to
  - Predict predicate-argument structure in polysynthetic languages
  - Surface complex morpho-syntactic features in isolating languages

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word
- Successful model will be able to
  - Predict predicate-argument structure in polysynthetic languages
  - Surface complex morpho-syntactic features in isolating languages
- The data will allow

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word
- Successful model will be able to
  - Predict predicate-argument structure in polysynthetic languages
  - Surface complex morpho-syntactic features in isolating languages
- The data will allow
  - More diverse cross-lingual studies

# Modelling Morpho-Syntax

- Models for analyso-parsing will ascribe
  - Features to any function – even if expressed periphrastically
  - Arc for every relation – even if between 2 parts of the same word
- Successful model will be able to
  - Predict predicate-argument structure in polysynthetic languages
  - Surface complex morpho-syntactic features in isolating languages
- The data will allow
  - More diverse cross-lingual studies
  - NLP tasks for low resource languages

# Thank you

Did I leave any time for questions/discussions?