Multilingual and Parallel Corpora
Parallel treebanking

Amir Zeldes

amir.zeldes@georgetown.edu

Parallel Treebanking

- We've seen fine-grained word alignment and sentence alignment
- Now we will look at full hierarchical alignment:
  - all sentence constituents
  - annotation of alignment types
- Work with the Stockholm TreeAligner:
  - http://kitt.cl.uzh.ch/kitt/treealigner/wiki/TreeAlignerDownload

Phrases

- Below the utterance level, we have individual words
- But words are grouped together in certain ways
- We call these groups phrase – but how much do we know about them?

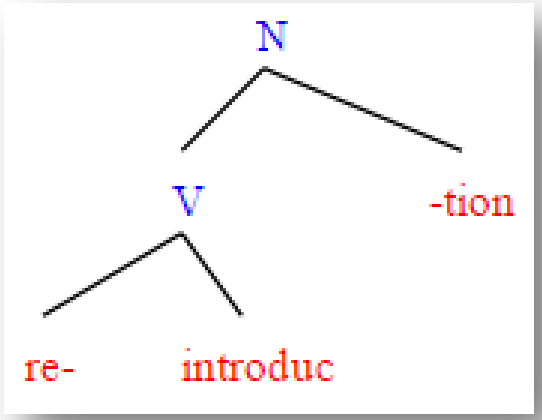
Phrases

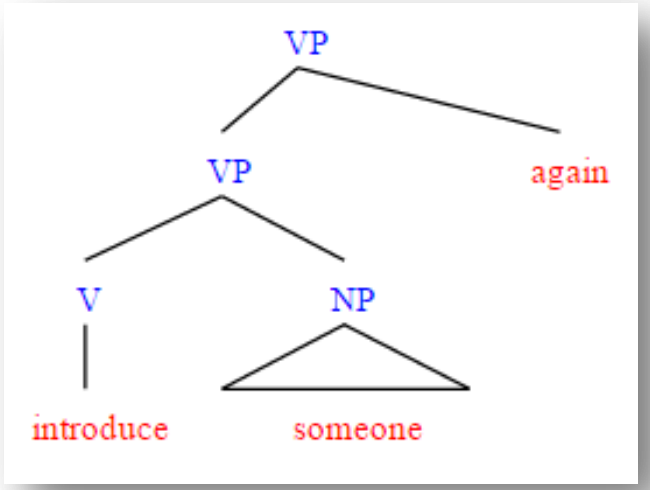
- What are phrases?
- How do we know they exist?
  - How can we tell what kinds there are?
  - Where they begin and end? Is every word a phrase? Are sentences phrases too?
- How are phrases structured?
  - What kind of structures do phrases form?
  - Are they always continuous?
  - (Binary?) trees? Graphs? Projective/non-projective, movement accounts…
  - How can we determine what's embedded in what?

Motivation for positing phrases

Let's start with an analogy to morphology:

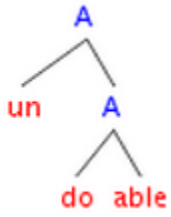
- Just as in morphology morphemes combine to form words, so in syntax: words form phrases
- Just like complex stems, combinations of words can combine again to form larger phrases

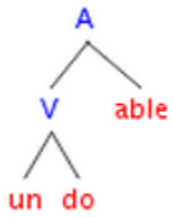




Motivation for positing phrases

- Like words, the hierarchical order of embedding influences meaning:





Motivation for positing phrases

- Like words, phrases are un-interruptible: (always?)

  Introd-re**-uction*

  introduced someone **I again*

- Behave like 'units' in a sentence, independent, complete
  - [The man with the dog]
  - [What [about [him]]]?
  - [[He] [came by]].

  - *The man with
  - *What about?
  - *Came by.

More formal tests

- There are three main tests for constituency:
  - Substitution
  - Permutation
  - Coordination
- The term constituent is preferred to phrase (more technical, less preconceptions)
- Caveat: Some use constituent to mean immediate constituent of a clause (e.g. argument, VP)

Constituency tests - Substitution

- Anything that can be replaced by a single element is a phrase or 'constituent'
- Theoretically: anything we already know is a phrase
- In practice: usually a pro-form

Constituency tests - Substitution

- Example: (Van Valin 2001:111)
  - The new teacher read a short book in the library.
  - She read a short book in the library
  - The new teacher read it in the library
  - The new teacher read it there
  - The new teacher read it in there
  - The new teacher did
  - * The new teacher read it in the there

Constituency tests - Substitution

- NB: did is a pro-VP, not a pro-verb
  - * The new teacher did a short book in the library
  - Is [read] a phrase?
- Not all PPs can be pronominalized:
  - The new teacher read it with the boy
  - The new teacher read it with him
  - *The new teacher read it therewith
- But still no such thing as:
  - glarf = [new teacher read] → The glarf a short book
  - wugged = [read a short] → The teacher wugged book

Constituency tests - Permutation

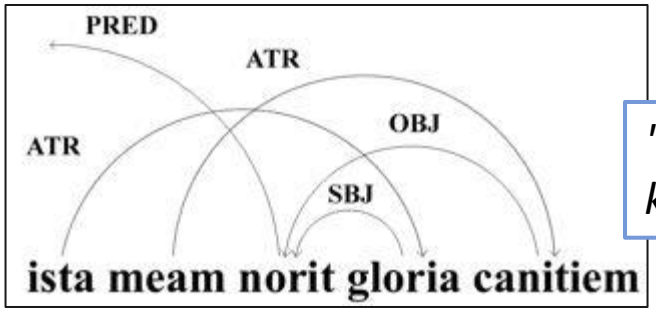
- We can rearrange constituents (in English):
  - In the library, the new teacher read a short book
  - *In, the new teacher read a short book the library
  - *A short the new teacher read book in the library
- But:
  - ? The library, the new teacher read a short book in
- And:
  - The teacher gave a book [to the student]
  - [Who] did the teacher give a book [to]?

Constituency tests - Permutation

- Permutation evidence for VPs:
  - The new teacher wanted [to read a short book in the library], and [read a short book in the library] she did
  - *The new teacher wanted to read a short book in the library, and [read] she did [a short book in the library]
  - ?The new teacher wanted to read a short book in the library, and [read a short book] she did [in the library]

Constituency tests - Permutation

- Other languages have no qualms about moving words around individually:
  - Latin: *magna cum laude* [[big with$_P$ praise]$_{NP}$]$_{PP}$
  - Greek: *t^{h}oas epi nēas* [[fast on$_P$ ships]$_{NP}$]$_{PP}$



"That glory shall know my old age"

Perseus Treebank, Propertius 1.8.46

Constituency tests - Coordination

- Only constituents may be coordinated:
  - [[on the table] and [under the chair]]
  - on [[the table] and [the desk]]
  - [[on] and [under]] the table
- Contrast:
  - *on the and under a table
  - *on the big and under a small table
  - *the happy and the angry boys (?? Van Valin)

How will constituents map across languages?

- Unlike ‘quick & dirty’ automatic alignment:
  - We can define constituents using linguistic knowledge in each language
  - Consider alignment at ***all*** possible levels
  - Classify types of alignment
- Most ambitious, fine-grained type of alignment so far!

Let's try this!

- We will work with three sentences from the UN Declaration of Human Rights:
  - http://research.ics.aalto.fi/cog/data/udhr/
- English constituent trees already available here:
  - https://corpling.uis.georgetown.edu/synannotri/
  - NetID <> 12345
- Find the corresponding sentences and construct your TL tree!

Step 1

- Paste a sentence into the interface
- Surround whole sentence with:
  - (ROOT …)

- Example:
  - **(ROOT** bla bla bla …**)**

Step 2

- Surround each token with brackets:
  - Write the word form at the right bracket
  - Add a part of speech on the left using:
    - N, V, DT, ADJ, ADV, PREP, PRO, COORD, CONJ, X
  - A token then looks like this: **(N** lizard**)**

- Example:
  - (ROOT (PRO she) (V is) (DT a) (N pianist))

Step 3

- Group phrases hierarchically using brackets and phrase labels written on the right:
  - **(NP** (DT the) (N pianist)**)**
  - Possible labels: S, NP, ADJP, ADVP, PP, VP, XP
  - You can use indentation with spaces to help readability
- Example:
  - (ROOT
    (NP (PRO she))
    (VP
    (V is)
    (NP (DT a) (N pianist))
)
)