Building Spoken Dependency Treebanks

Building Spoken Dependency Treebanks

Kim Gerdes
LPP
Sorbonne Nouvelle

Outline

- About me: University and Projects
- Spoken Language:
  - Interest in Spoken Language
  - Preparation before syntax:
    - Minimal and maximal units
    - Macrosyntactic annotation
  - Possible workflows
- Annotation
    - Some problems
    - The tools
    - Bootstrapping

Personal Presentation

Personal Presentation

- Kim Gerdes 凱德金
- Formal Linguist
- Master in Pure Mathematics, PhD in NLP
  - Syntax and Natural Language Generation
- Associate Professor at the Sorbonne Nouvelle, Paris
  - Sorbonne?

Sorbonne!

- University of Bologna (1088)
- **University of Paris** (teach. mid-11th century, recogn. **1150**)
- University of Oxford (teach. 1096, recogn. **1167**)
- University of Modena (1175)
- University of Palencia (1208)
- University of Cambridge (**1209**)
- University of Salamanca (1218)
- University of Montpellier (1220)
- ...

Personal Presentation

- Kim Gerdes 凱德金
- Formal Linguist
- Master in Pure Mathematics, PhD in NLP
- **Associate Professor at the Sorbonne Nouvelle, Paris**
  - **Formal Syntax**
    - **Dependency Linguistics:**
      - **Co-organizer of the *International Conferences on Dependency Linguistics*: Depling. This year in Pisa**
  - **Corpus Linguistics**
  - **Free Resource Development**
    - **Treebanks, Software**
    - **Belief:**
      - **Publicly sponsored resources should be free, just like publicly sponsored publications can be used (fairly) by anyone.**

Machine translation preprocessing

- 2011/2012 Visiting Scholar at the Chinese Academy of Sciences 国科学院
  - With Zong Chengqing 宗成庆 , Peking
- Integration of
  - morphological tools in a preprocessing chain
  - German specific named entity recognition
- Bilingual Alignment of similar texts
- Tool development for usage of bilingual online resources

Personal Presentation

- **Work with cloem.com**
  - **Natural Language Generation in the domain of Intellectual Property**
  - **Help for creation of patents**
  - **Artificial Creativity: The last frontier before the *Technological Singularity***

Projects

- 2008-2012: Rhapsodie
  - 30 thousand words of Spoken French
- 2012-2016: Orféo
  - 3,5 million words of Spoken French + 6,5 million words of Written French
- 2016-2017: Procore Hong Kong
  - Parallel treebank of spoken Cantonese with Mandarin translation
- Since 2017:
  - NaijaSyncor
  - Profitérole: Old French

French Spoken treebanks

- ANR Rhapsodie 2008-2012 (directed by Anne Lacheret)
  - 33 000 words = 3hours
  - 57 samples, 5 min long, from various genres
  - annotations : prosody, microsyntax, macrosyntax
- ANR Orféo 2013-2017 (directed by J.-Marie Debaisieux)
  - 3,5M words spoken (350h) + 6,5M words written
  - microsyntax : 150 000 words manually annoted and the rest automatically parsed
  - query platforme

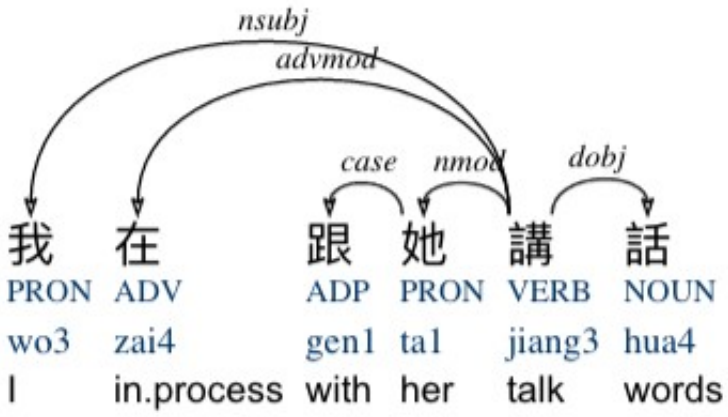
Other treebank projects

Remote involvement with “Universal Dependencies” international community: 100 syntactic treebanks of written texts in 50 languages, using the same tag set.

- Naija, Old French, Orféo have or will have a UD version.
- Very critical about the syntactic choices
  - Confusion of semantics and syntax
  - Confusion of category and function
  - Not well adapted to spoken language phenomena

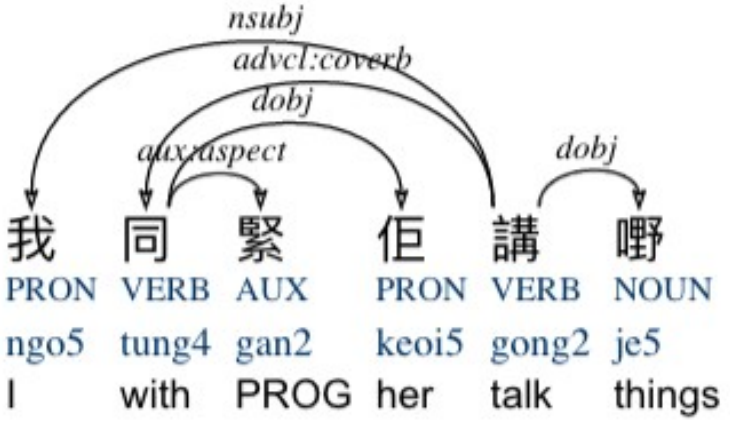
Chinese-Cantonese Bilingual treebank of Cantonese TV dramas and Youtube clips

- Hubert Curien Partnership with John Lee, Hong Kong City University
- Goal: create a *dependency treebank* for Mandarin Chinese to support comparative linguistic studies, e.g., Cantonese vs. Mandarin



Mandarin

‘I am talking with her’



Cantonese

Spoken Language

Spoken Language

- Recognized facts in most frameworks:
  - Data-driven / empirical approaches can give new insights
  - Spoken language is the « purer » form of language that gives deeper insights
- Why so few corpora are spoken corpora?
  - History of Syntax as prescriptive grammar
  - Need in NLP to process written text
  - Easy to capture large amounts of written data on the Web

Spoken Language Treebanks

- Treebanks: corpora with syntactic structures allow:
  - Empirical syntax
    - Direct access to examples
    - Quantitative measures
  - Training of parsers
- Even less spoken treebanks:
  - Constituant structure and analysis of corrected parts:
    - Switchboard Corpus of Penn Treebank
    - (Meteer et al. 1995)
    - CHRISTINE (SUSANNE analytic scheme; BNC)
    - ICE-GB
    - VERBMOBIL (English, German, Japanese)
  - Dependency, but also only “corrected” parts are analyzed
    - CGN (Spoken Dutch Corpus; Schuurman et al. 2004)

So why is dependency hip?

- recent years of NLP: steady rise of the importance of functional dependency annotations. Reasons for this might be:
    1. the consideration of languages having freer word order than English and French
        - for which the phrase structure grammars, predominant in corpus linguistics up to then, prove to be insufficient.
    - Also spoken language tends to have more non-contiguous structures, afterthoughts, inserts, etc.

- *Basic assumption of X-bar structures:*
    - *Coincidence of linear groupings and head-daughter relations*
    - *If not: movement*
- *So, dependency is easier and less controversial*

So why is dependency hip?

2. a more general change of linguistics paradigms, increasingly separating functional and constituent structures
3. the growing interest in the lexical subcategorization frames of words, which naturally leads to functional descriptions of grammar;
4. the increasing capacities of the automatic language tools
   - surpassing simple feature enriched context free grammars
   - obtain *deeper* structure, closer to semantics: interesting for analysis and generation

What to study with a treebank of spoken language?

- Syntactic sentence structure
- Prosodic structure
- Relation between the two
- Comparison of two languages
- Typological classification

→ different annotation choices

Sampling

- What is representative?
- What can we record with sufficient quality?
- We want natural speech.
  - Is read news spoken language?
  - Is TV drama natural language?
  - Where to put the mic?
- Ethical questions:
  - Tell the recorded people before or after the recording?

Transcription

- Surprisingly hard question

  Transcribing is analyzing

- Not only for learners' corpora
- Rhapsodie: >50 pages transcription guide
  - Basic idea:
    - No phonetic transcriptions. all transcriptions are made with words of the French language
    - If it is possible that the speaker actually spoke grammatically correct (i.e. according to the written grammar), we transcribe that
      - Example: *on **n'**est pas partis.*

Transcription

- Linguistic question:
  - Is there one language that is written and spoken? - even if we clearly observe different rules?
- Political questions:
  - How to transcribe Naija?

    Naija = Nigerian pidgin spoken by ~ 100 million people, increasingly natives

    is this English, with a strange pronunciation and a strange syntax?
  - How to transcribe Cantonese?

    Is this a dialect or a language?

    Do we use traditional or simplified characters?
    - In particular for a parallel corpus, it could be useful to choose either system for both sides

Minimal and maximal units

- What are the nodes of the syntactic tree?
  - Minimal units

    → tokenization
- What are the “sentences”?
  - Maximal units
    - How long are relations syntactic?
    - When do relations become discursive?

Minimal units

- Chinese:
  - Characters?
  - Words? What are words?
  - Usually: automatic lexicon-based tokenization
    - Advantage: first step of an automatic analysis sequence
    - Disadvantage: often not syntactically coherent.
  - Our treebanks: manual segmentation:
    - Disadvantage: no corresponding automatic tokenizer → we can train a statistical parser, but we can't automatically obtain the corresponding segmentation
    - Work in progress…

Minimal units

- Alphabetic languages:
  - Sequence of letters?
    - he's → he + 's
    - I'a *'has it'* → I' + a *it* + *has*
  - Lexicon?
    - Lexically non-compositional structure
      - *on top of* → one word?
        → slippery slope
    - Names? Named entities?
      - *The representative of the Embassy of Burkina Faso* → how many words?
      - Recursiveness: *the* ***United States Department*** *of Housing and* ***Urban Development***
    - MWE often ambiguous without context
      - *He told the ministry of foreign affairs*

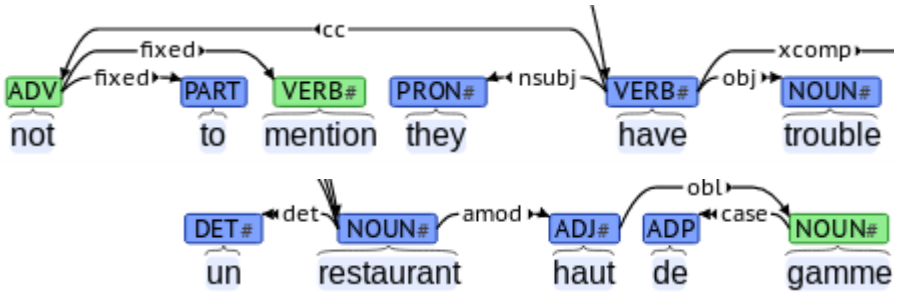
Minimal units

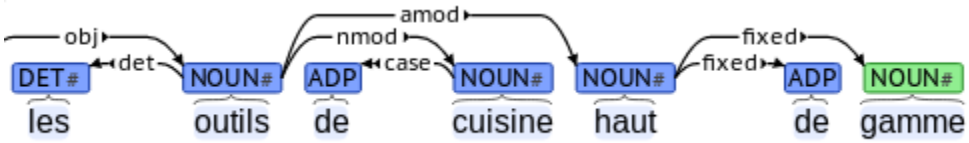
- Basically two choices:
  - Character-based rules
  - Lexicon-based rules
- Universal Dependencies: Character-based rules (except contractions that are undone: *aux*, *besame*)
  - Needs no lexicon
  - Needs specific relations that describe MWE
- French Treebank, Rhapsodie, Penn Treebank and most other treebanks: lexicon-based tokenization
  - Needs lexicon (criteria?)
  - Needs only real syntactic relations

Minimal units

- Endocentric vs exocentric MWE
  - Head has the same POS as the whole?

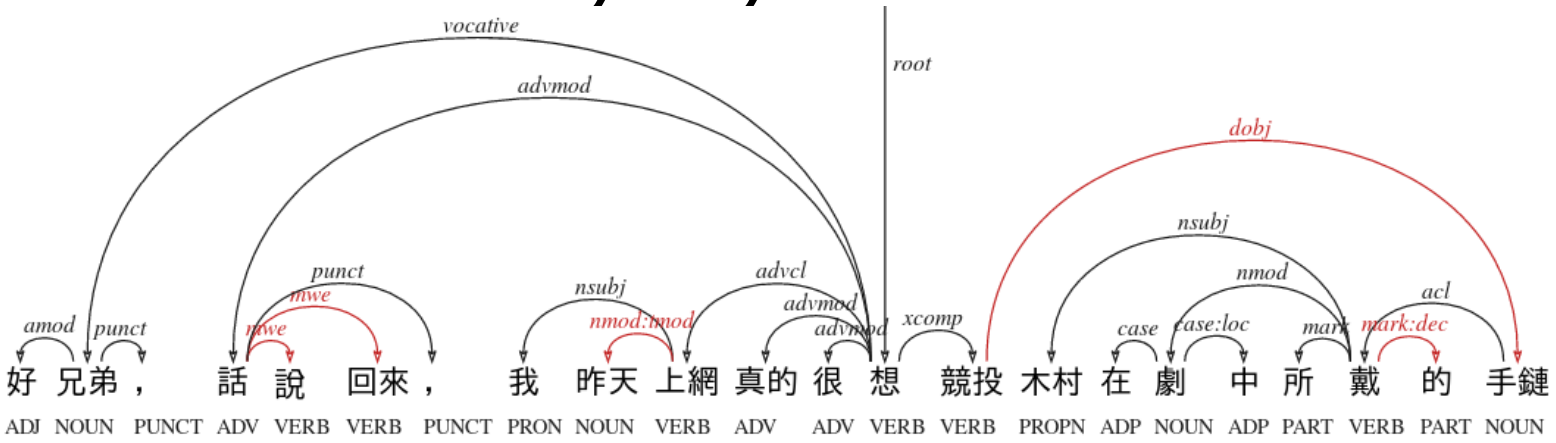
    → no problem
- Examples of MWE:
  - Not to mention
  - Contrary to
  - *Haut de gamme ‘top notch”: very top notch*



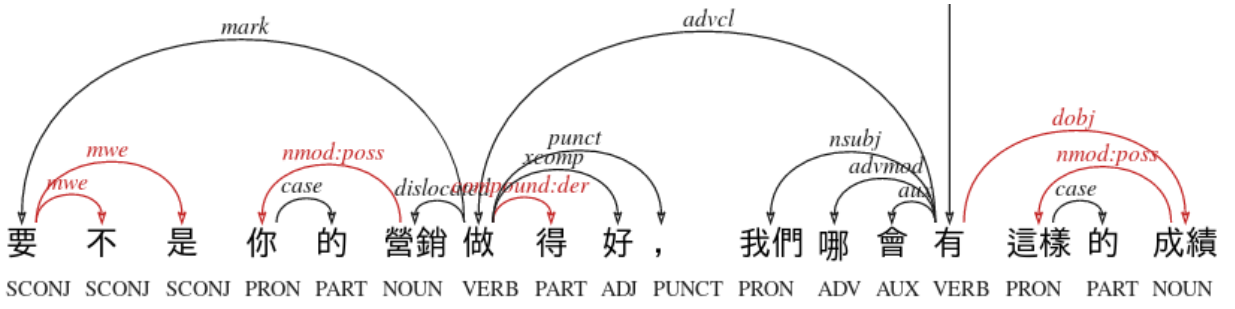


Minimal units

- Examples of MWE in Mandarin:
  - huàshuō huílái *anyway*

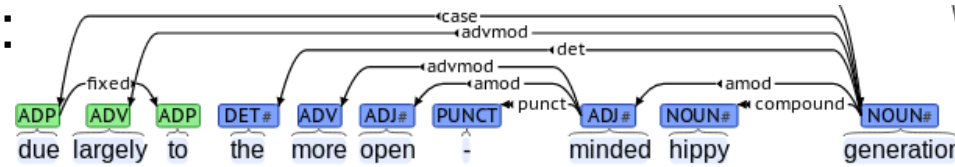


  - Yào bùshì *if it were not for or without*.



Minimal units

- Character-based tokenization rules are preferable because:
  - Deciding on the MWE status of a sequence is a syntactic question
  - Parsers (statistical or rule-based) are specialized in context-dependent analysis
  - The tokenization remains stable and thus comparable under different annotation schemes
  - MWE can be discontinuous:



- UD develops contractions → text becomes unreadable, and again, syntactic decisions have to be made before we have access to the syntactic structure
  - *Il mange souvent **de** belles choses.*

Maximal units

- Transcription does not include punctuation
- No prosody-based and word-based tool does a good segmentation in sentences.
  - Problem for automatic systems, work in progress…
- Cantonese has punctuation in the Mandarin subtitles – we mostly used that
- For Rhapsodie, Orféo, and Naija we apply a manual "macrosyntactic" pre-annotation:
  - Segmentation into Illocutionary Units
  - Separate the sentence core from "extraposed" elements
  - Mark direct speech
  - Mark piles: disfluency, elaboration, coordination

  → Think of it as some kind of specific punctuation

Segmentation

- Maximal syntactic unit = “sentence”
  = dependency domain + non autonomous segments

- **S1:** so i see that you're from Hartland Michigan this is right up the road
  S2: mhm like forty minutes from here

- **S1:** so i see that you're from Hartland Michigan

  this is right up the road
  S2: "mhm" like forty minutes from here

Disfluencies (or speech repairs)

- Original:

  Show me flights from Boston on uh from Denver on Monday
- Annotated (Shriberg 1994, Levelt 1983):

  Show me flights {from Boston on | "uh" from Denver on Monday }
  - Reparandum: from Boston on
  - | = interruption point
  - Interregnum: uh
  - Repair: from Denver on Monday
- Cleaned: Show me flights from Denver on Monday

Cleaning?

donc pour essayer un petit peu de sortir cette personne de la misère
(car c'est vraiment un petit peu semblable aux Misérables de Victor Hugo)
nous essayons tant bien que mal de lui faire comprendre que sa cabane
dans quelques années (entre parenthèses elle a 79 ans)
quand elle aura des difficultés (ce qu'on espère pas)
des difficultés à se déplacer ou à évoluer
(c'est-à-dire qu'il y a énormément d'escaliers à monter pour arriver à sa cabane)
donc le jour où elle ne pourra plus se déplacer
ou qu'elle sera malade un petit peu plus sévèrement
on essaye de lui faire comprendre qu'elle ne pourra plus vivre dans cette cabane

example studied by Blanche-Benveniste

Cleaning?

so to try a little bit to get this person out of misery
(because it's really a bit similar to the Misérables by Victor Hugo)
we try the best we can to make her understand that **her shack**
in a few years (let me add that she is 79 years old)
when she will have difficulties (which we don't hope for)
difficulties moving or turning around
(meaning that there are lots of staircases to go up in order to get to her shack)
so the day she won't be able to move around
or she will fall sick a little bit more seriously
we try to make her understand that she won't be able to live in this shack

translated from French, example studied by Blanche-Benveniste

Cleaning?

- We don’t want to clean the disfluencies because:
  1. **We are interested in their prosodic contours (Rhapsodie project)**
  2. **We think there is a continuum up to constructions that cannot and must not be cleaned (Aix school)**

From disfluency to coordination

- She is a linguist, maybe a computer scientist
- She is a linguist uh maybe a computer scientist
- She is a linguist or maybe a computer scientist
- She is a linguist and maybe a computer scientist
- She is {a linguist | maybe a computer scientist }

maybe is a paradigmatizing adverb (Nølke 1983, Pietrandrea 2009)
She is maybe a computer scientist
*She is and a computer scientist

Modeling

- Heeman, McMillin & Yaruss 2006
  (annotation of stutterers' productions)

  She is a linguist maybe a computer scientist

  - disfluency = **backtracking**
- Grid analysis (Blanche-Benveniste et al. 1979)

  She is a linguist
  maybe a computer scientist
  - disfluency and coordination = **paradigmatic piles**
  - Layers are both in syntagmatic and paradigmatic relations

Rhapsodie's encoding

- She is {a linguist | ^and a computer scientist }
- She is { a linguist
  | ^and a computer scientist }
- She is a linguist
  and a computer scientist

- | = junction point (= interruption point for disfluency)
  { = backtrack point
  ^ = pile marker (conjunction of coordination)
  } = ???

} in disfluencies

- okay so what what changed your mind and what has it been changed to
- okay so {what
  | what changed your mind
  | ^and what has it been changed to }
- okay so {{what | what } changed your mind
  | ^and what has it been changed to }
- okay so {what &
  | what changed your mind
  | ^and what has it been changed to }
- & = incomplete syntactic unit

} in coordination

- For coordination, } can indicate the scope
  - a boy and a girl I met yesterday
  - {a boy | and a girl } I met yesterday
  - {a boy | and a girl I met yesterday }

Macrosyntactic annotation

and {i | i} have lots of other interests
{like "um”
| that are a little bit more like }
{paleontology
| ^or astronomy
| ^or international religion
| ^or "uh" not religion
| international relations
| ^so those things
{i wanna &
| i think i'm gonna concentrate more on } }

(Micase 1:18)

Maximal units

- Only after the macrosyntactic markup, the actual dependency annotation starts.
  - Rhapsodie: dependency links only inside of each unit (left-dislocated, core, right-dislocated)
  - Orféo: One tree per illocutionary unit

    *a) **This situation and all**, my job has become unbearable*

    *b) **Since June**, my job has become unbearable*

    *a) Peter is not here, **because I haven't seen his car in the parking lot***

    *b) Peter is not here, **because he is really sick this week***

    - One or two syntactic units?

macrosyntax

- pure peripheric element vs core element
  - *to my **knowledge** < nothing is possible*
  - *with **that** <+ nothing is possible*
- No distinction in UD:
  - both are *obl(ique)*?
  - Both are *dislocated*?
- Create subclasses of functional relations:
  - *dislocated:periph vs. dislocated:obl*
  - *periph:dislocated vs. obl:dislocated*
- advcl
  - *If I am tired <+ I go to bed*
  - *If I understand what you say < you are ok*

Graft

- The speaker grafts a periphrastic sentence in a syntactic position where an NP is required (Deulofeu 1999, 2007)
- Yesterday i met **[i think his name is <u>John</u>]**

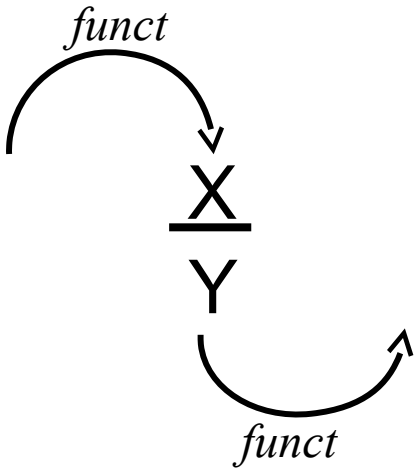
Graft

it's being done in the whole field of
[is there a common ancestor
| or did a humanoid species
{spring up
| ^or [S2: mhm] **exist } in various places**
{in the world [S2: mhm]
| not just in Africa
| ^but also in Asia
| ^and maybe also in southern Europe [S2: mhm] **}]**

(Micase 2:58)

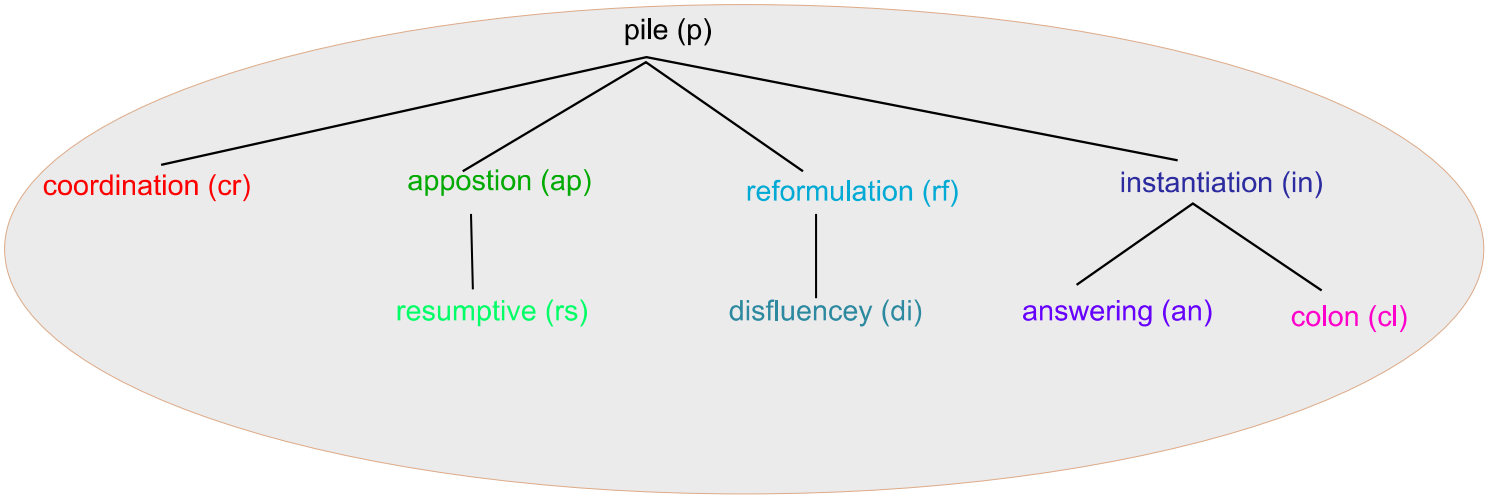
Paradigmatic piles

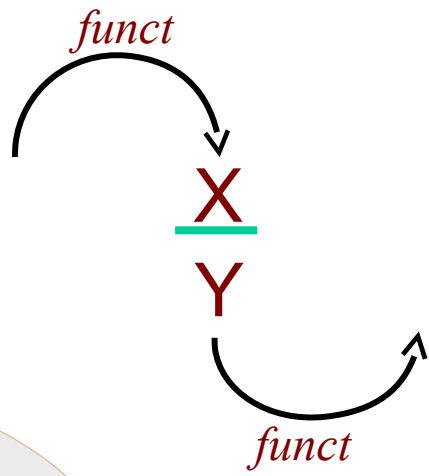
- a segment Y of an utterance *piles up* with a previous segment X if Y fills the same syntactic position as X
- Y can be a (voluntary or involuntary) reformulation of X, Y can instantiate X or Y can be added to X in a coordination



Paradigmatic piles

- uses of piles:
  - disfluency, reformulation, apposition, instantiation, colon-effect, question-answer relationships, coordinations and intensification





Colon effect

- I bought many things paper glue and so on
- Written:
  I bought many things: paper, glue, and so on.
- Annotation:
  I bought {many things | paper | glue | ^and so on }
- Hierarchization:
  I bought {many things | {paper | glue | ^and so on } }
- Typing:
  I bought {many things $|_{:}$ paper $|_{+}$ glue $|_{+}$ ^and so on }

Colon effect

it’s being done in the whole field of
[is there a common ancestor
| ^or did a humanoid species
{spring up
| ^or [S2: mhm] exist } in various places
{in the world [S2: mhm]
| not just in Africa
| ^but also in Asia
| ^and maybe also in southern Europe [S2: mhm] }]

(Micase 2:58)

Instanciation

- Colon-effect can produce long-distance piling:
  - I would say that you gave {something more } to the woman {* arms of persuasion }
  - well in fact there are {quite a few things } that contribute already uh {* the atmosphere in the shop | …}

Question-answer relationships

- S1: {when } do you plan to do that?
  S2: {* today}
- Instanciation
  - of an interrogative pronoun
  - of a thing-term (colon-effect)

Criss-cross

- One criticized the the newspaper of I think it was the Provençal one criticized it in relation to or the Méridional in relation to the death of what was his name not Coluche the other one

(translated from French, Blanche-Benveniste 1990)

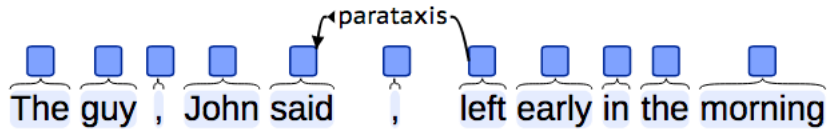
- {one criticized
  {the
  | the newspaper of
  | [I think it was {the Provençal |}] }
  | one criticized it
  {in relation to
  {| ^or the Méridional}
  | in relation to} the death of
  [{what|} was his name
  {| ^not Coluche
  |the other one}] }

Intensification

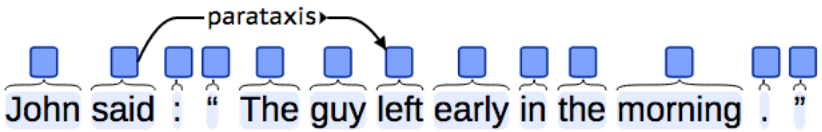
- This is a very very serious question
- Make it quickly quickly quickly
- I gave examples examples examples

parataxis

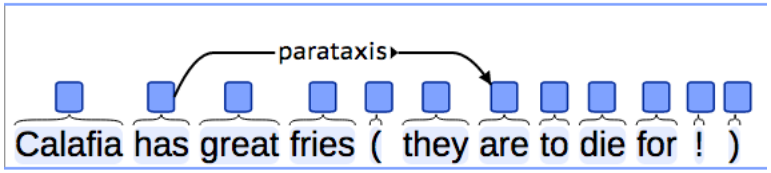
- parataxis relation in UD covers:
  - parataxis:insert



  - parataxis:ccomp or ccomp:direct



  - parataxis:parenthesis



Howto

Rhapsodie workflow

- Pivot:
  - transcription (one tier for each speaker)
  - sound-transcription alignment
- Syntax:
  - linearization of the transcription (including two additional tags: speakers, overlaps)
  - macrosyntactic annotation
  - automatic segmentation (into microsyntactic units)
  - automatic dependency parsing (with a rule-based parser for written French)
  - manual correction
- Independently: Prosody
  - Problem: prosodists and syntacticians find errors in the description independently.

Orféo workflow

- transcription (one tier for each speaker)
- sound-transcription alignment
- (weak) macrosyntactic annotation (for each speaker)
- automatic segmentation into “sentences”
- linearization (no speech turn inside a “sentence”)
- bootstrapping:
  - conversion of the Rhapsodie treebank (including macrosyntax) for priming
  - training of a parser
  - manual correction (of parts of the corpus)

Naija workflow

- transcription (one tier for each speaker)
- sound-transcription alignment
- macrosyntactic annotation (for each speaker)
- linearization (no speech turn inside a “sentence”)
- Manual POS tagging (lexicon-based)
- 1st bootstrapping step:
  - English Glossing of tokens
  - Parsing with Mate parser trained on UD English
  - manual correction
- 2e bootstrapping step:
  - Training on first corrected sample set of Naija
  - Manual correction

Cantonese workflow

- Rule-based transformation of parser trained on Chinese Penn treebank into Orféo-type annotation
- Training of parser
- Parsing of Mandarin sub-titles
- Rule-based transformation into UD v1.2
- Manual correction
- Development of annotation guide
  - Classifiers caused the biggest trouble
  - Proposed adding clf=classifier relation to UD
  - Accepted for UD v2 (février 2017)
- Rule-based transformation into UD v2
- Manual correction
- Manual annotation of Cantonese
- In parallel:
  - Rule-based induction of Cantonese treebank by word-similarity
  - Manual correction of Cantonese
- Ongoing project…

Tools

- Basically two types of tools:
  - For the creation of the treebanks:
    - Taggers, parsers
    - Manual multi-annotator correction and validation
  - For the exploitation of the treebanks:
    - Concordancer
      - Simple or with query language for specific searches
    - Global statistical frequency and collocation measures
- The tools cannot be clearly separated:
  - Need for searches and measures for error mining the treebank.

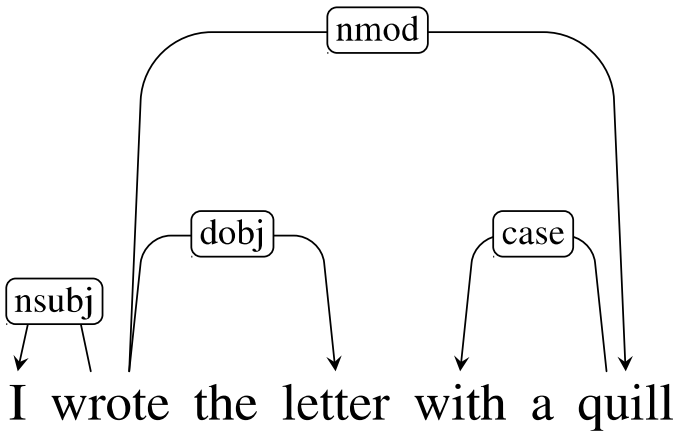
Tools

- Arborator: manual correction of dependency annotation
- MATE parser (Bernd Bohnet)
- Orféo platform (Orféo engineer)
  - consultation by samples
  - solr request
  - ANNIS QL (treebank query language)
- UD platform

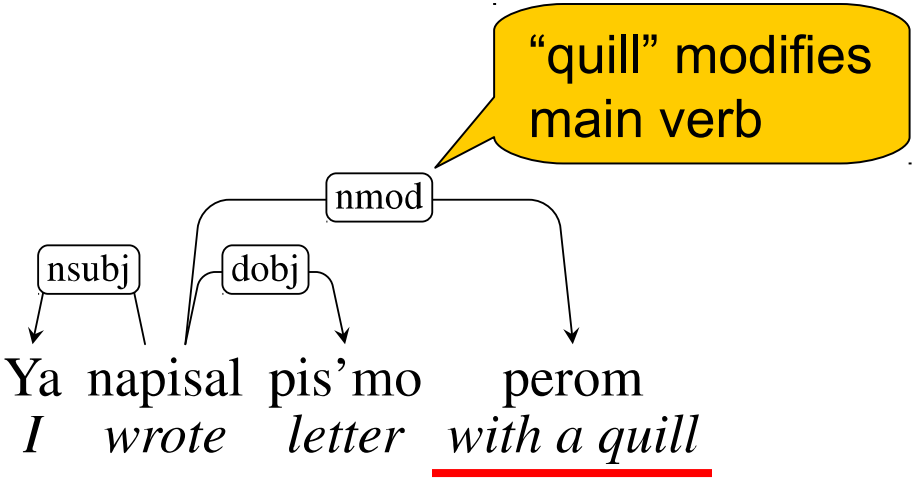
Annotations

Universal Dependencies (UD)

- Example: PP (cont'd)
  - Many languages use case rather than preposition
  - UD annotation yields more similar structure



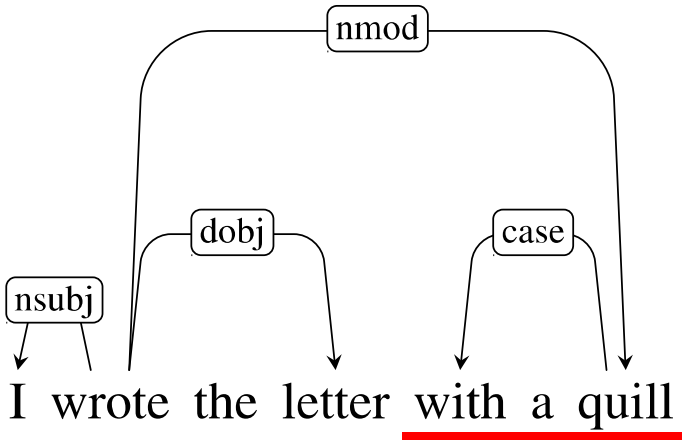
English



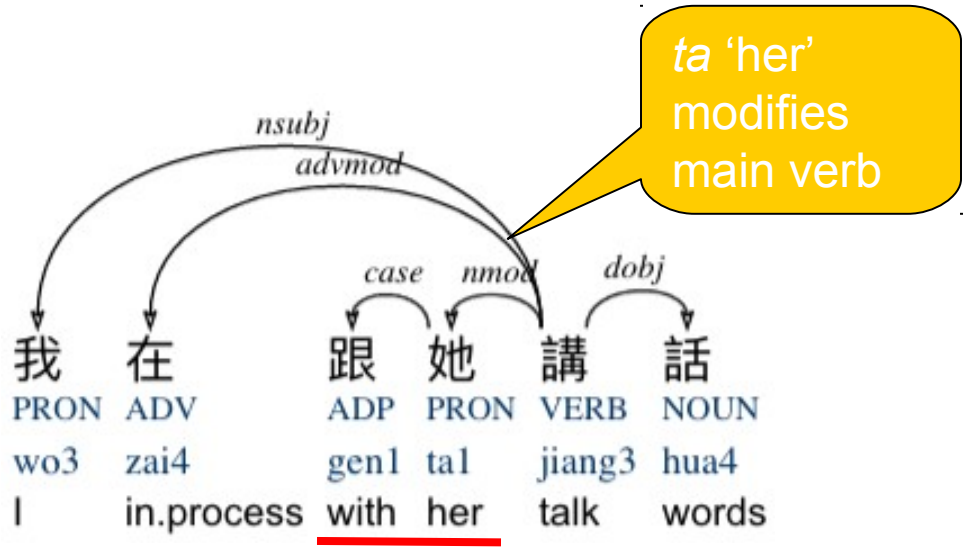
Russian

Universal Dependencies (UD)

- Example: PP (cont'd)
  - Mandarin coverbs function like prepositions
    - Cf. Preposition stranding (Huang 1982, McCawley 1992)



English

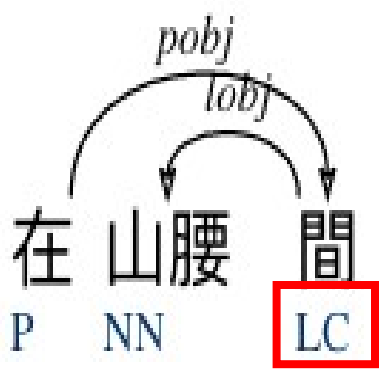


Mandarin

'I am talking with her'

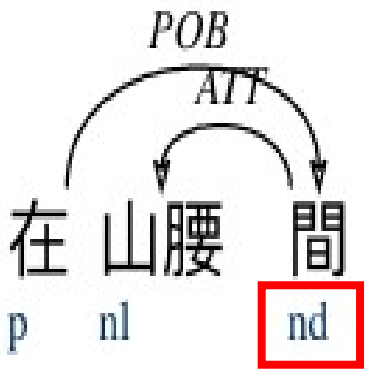
Localizers

- A localizer follows a noun to indicate spatial information
  - e.g., *shang* ‘above’, *jian* ‘middle’
  - Often paired with locational preposition *zai*
  - Given special tag in Stanford and CDT



Stanford

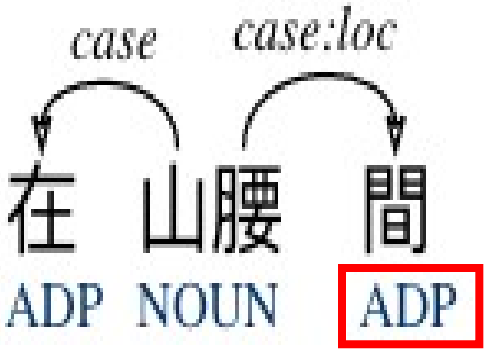
CDT



‘on the mountainside’

Localizers

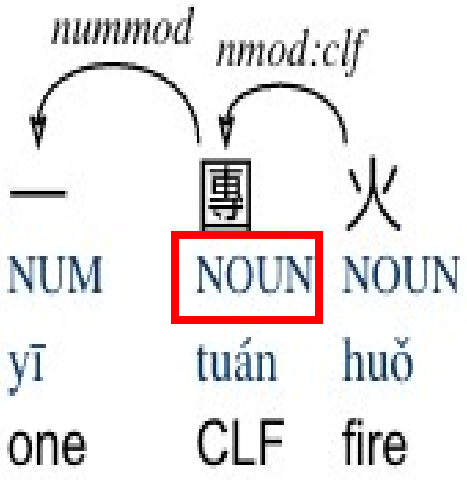
- Mandarin UD treats localizers as adpositions (`ADP`)
  - Unlike nouns, localizers cannot be used on their own
  - Postpositional analysis supported by some research (Peyraube, 1980; Ernst, 1988)
    - Cf. cross-linguistic observations (Djamouri, Waltraud, & Whitman, 2013; Waltraud, 2015)



'on the mountainside'

Classifiers

- In Mandarin, classifiers are often obligatory with numerals and determiners
  - Tagged in Stanford as ‘measure word’ (M)
  - Tagged in CDT as ‘quantity’ (q)



‘a ball of fire’

Universal to do what?

- UD makes typologically different languages look similar.
- Lexical items appear higher in the tree
  - Useful for simple extraction of semantic structures
  - Possibly for translation studies of the lexicon

Universal to do what?

- But: typological differences are hidden and cannot be measured
  - *Note that there are words that may be traditionally called numerals in some languages (e.g. Czech) but they are treated as adjectives in our universal tagging scheme. In particular, ordinal numerals ... behave both morphologically and syntactically as adjectives and are tagged ADJ.*

第一 dì yī ???

Hands on

- The Arborator
  - arborator.ilpga.fr
  - Has been used for French, English, Occitan, Chinese, Swedish, Naija etc.
  - At Paris 3 Sorbonne Nouvelle, Paris Ouest Nanterre, Lorraine University, at Xi'an Jiatong, University of Uppsala, City University of Hong Kong, Indiana State University, Georgetown University, University of California – Davis
  - Also used for teaching dependency
    - Specific teaching modes

唔該！

Kim Gerdes

kim@gerdes.fr

http://gerdes.fr

Other tools

ORFÉO

Texte | Rechercher | Chercher

Limiter votre recherche

Corpus

Language

Number of speakers

2	460
1	134
3	121
4	35
5	27
6	21
8	10
9	10
7	8
10	4

Text type

entretien	334
conversation	246
narration	118
présentation	51

Bienvenue!

Bienvenue sur la page de recherche de texte du projet Orféo

ORFÉO

English | français

Historique de recherche

Text | facile | Chercher

Voir extraits de texte | Voir concordancier

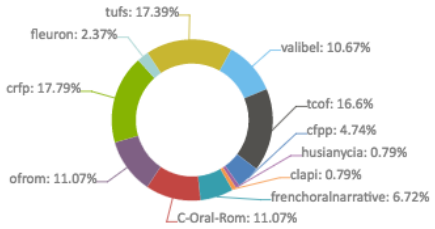
Limiter votre recherche

- Corpus
- Language
- Number of speakers
- Text type

Vous avez demandé : facile

Accueil

253 résultats



5 mots précédents | 5 mots suivants

№	Nom de fichier	Type	Corpus	Contexte gauche	Résultat	Contexte droit	Liens
1	Bloch_052-...		frenchoralnarrative	pas et c' est pas	facile	à expliquer	
2	ffammn20		C-Oral-Rom	la vallée de Chamonix très	facile	à atteindre en fait il	
3	ffammn20		C-Oral-Rom	un chemin un chemin très	facile	un petit peu escarpé et	
4	ffammn20		C-Oral-Rom	qu' il permet une descente	facile	et rapide ce jour-là notre	
5	Kiss_202i-...		frenchoralnarrative	drôle c' est vraiment pas	facile	à vivre le quatrième se	
6	Nataf_041-...		frenchoralnarrative	demandait vous avez un métier	facile	pour mon fils un métier	
7	Nataf_041-...		frenchoralnarrative	pour mon fils un métier	facile	pour mon fils vous n'	
8	Nataf_041-...		frenchoralnarrative	n' auriez pas un métier	facile	pour mon fils un métier	
9	Nataf_041-...		frenchoralnarrative	pour mon fils un métier	facile	pour mon fils et partout	
10	Nataf_041-...		frenchoralnarrative	disait mais allez un métier	facile	ça n' existe pas ici	

Affichage des lignes 1 à 10 sur 253 lignes au total | 10 | lignes par page

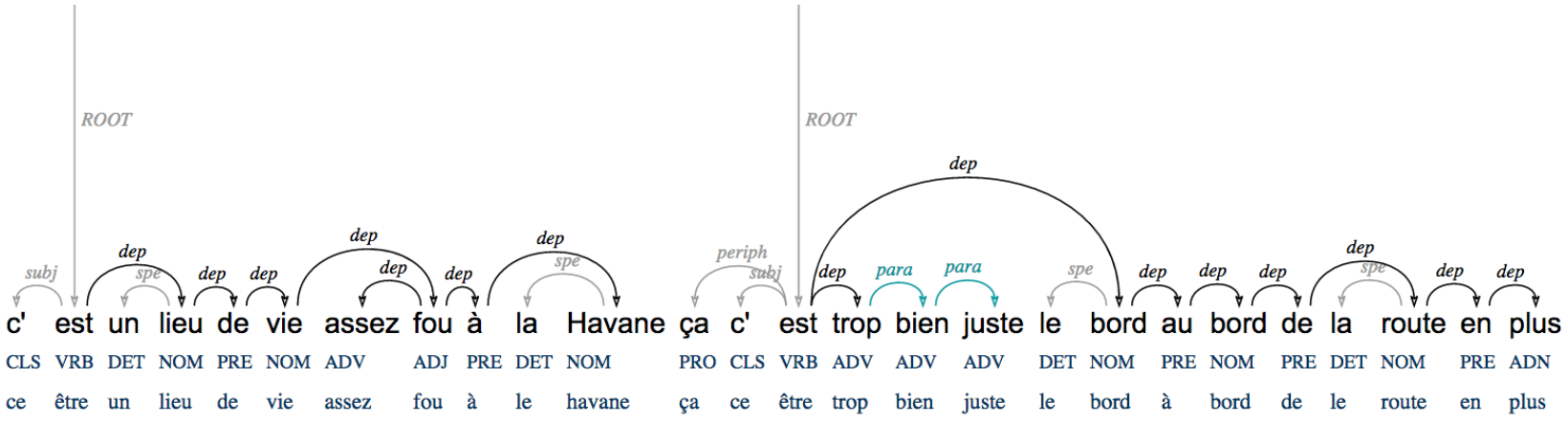
‹ 1 2 3 4 5 ... 26 ›

ffamdl10

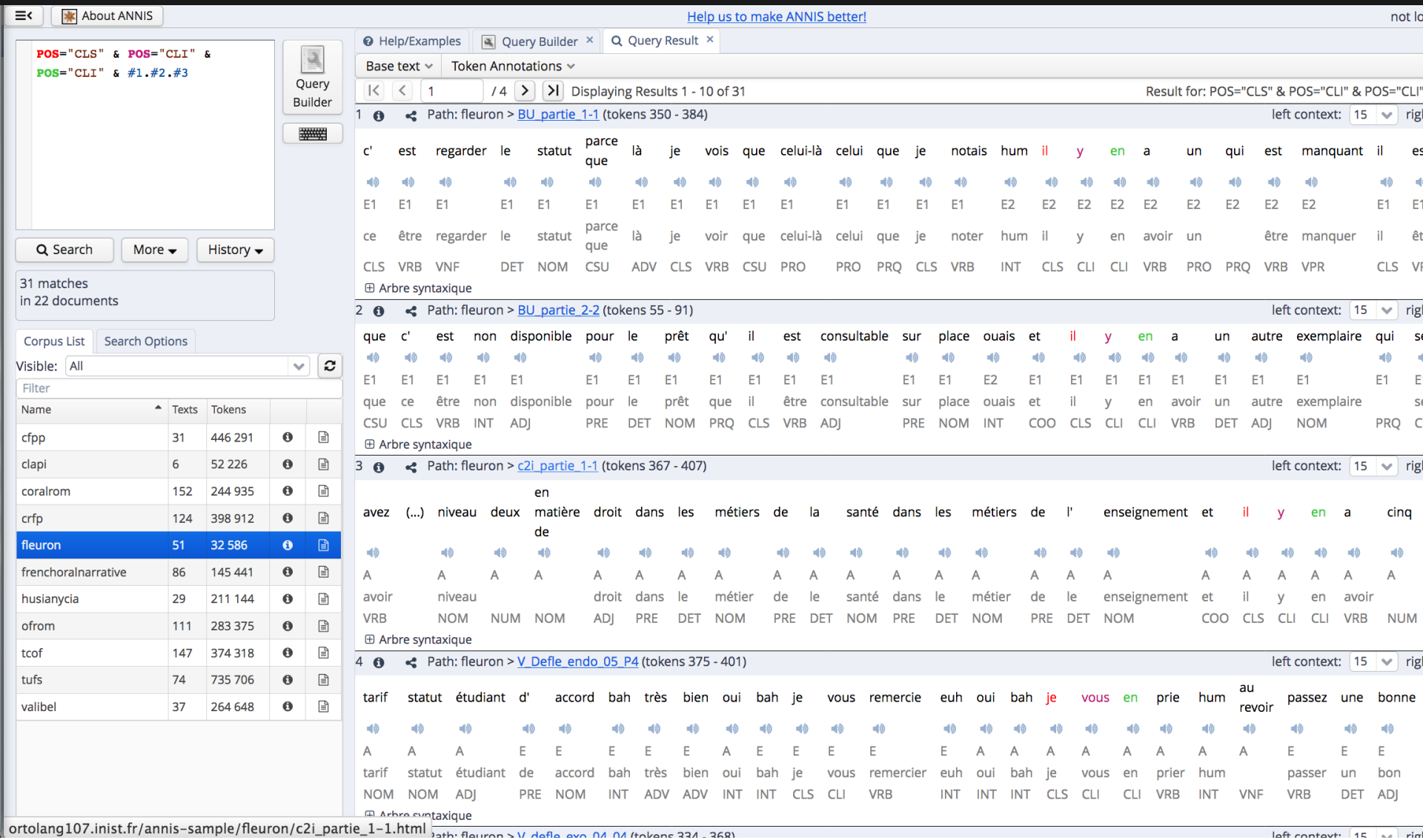
Le locuteur raconte son voyage à Cuba
(un échantillon dans le corpus **C-Oral-Rom**)

- Corpus C-Oral-Rom ?
- Métadonnées : general ?
- Métadonnées : locuteur DAV ?
- Métadonnées : locuteur DAM ?
- Texte et audio ?
- Arbres syntaxiques ?

1: c' est un lieu de vie assez fou à la Havane ça c' est trop bien juste le bord au bord de la route en plus



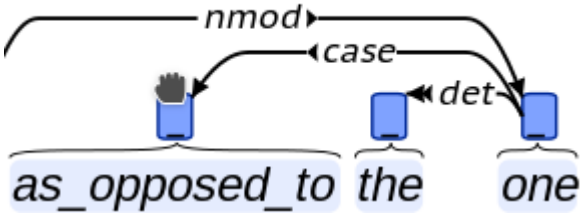
ANNIS Query Language (Zeldes et al. 2009)



Universal problems

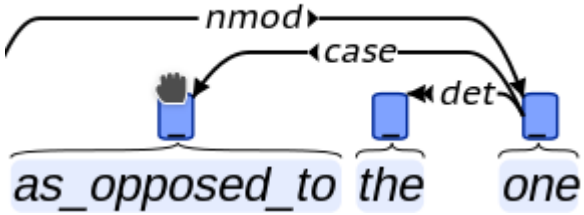
MWE “as opposed to” choice A: one token

preposition as one token



MWE “as opposed to” choice A: one token

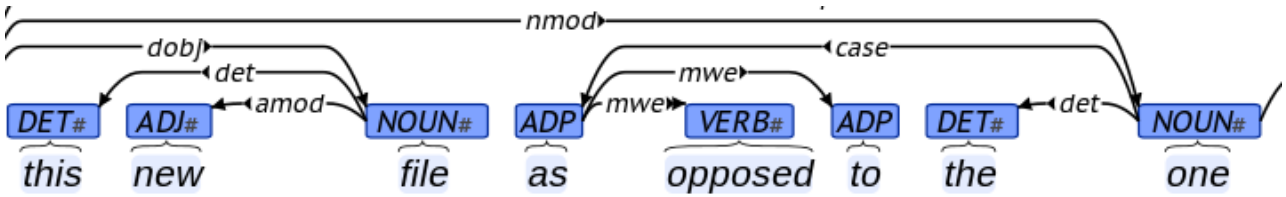
preposition as one token



– Pre-annotation needs large lexicon

+minimality, but –concision

MWE “as opposed to” choice B: UD

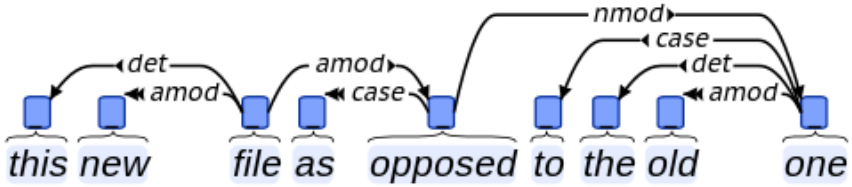


special link *mwe*

Annotator needs access to lexicon!

–concision

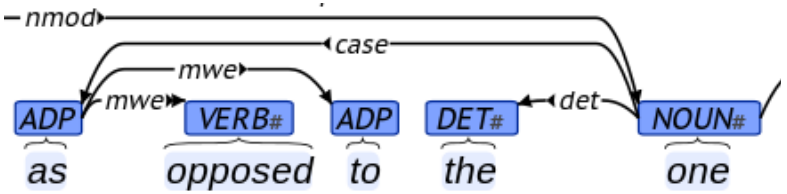
MWE “as opposed to” choice C: compositional



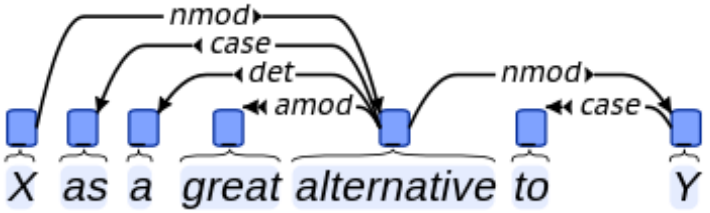
No distinction of MWE and free constructions

MWE “as opposed to” continuum

- UD: special link:

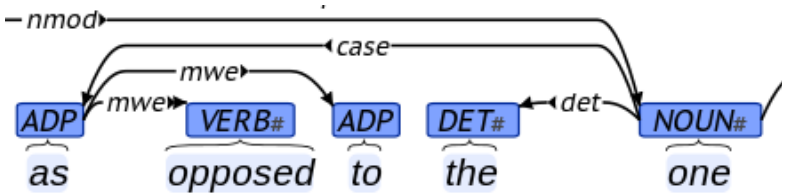


- UD: Compositional:



MWE “as opposed to” continuum

- UD: Special relation:



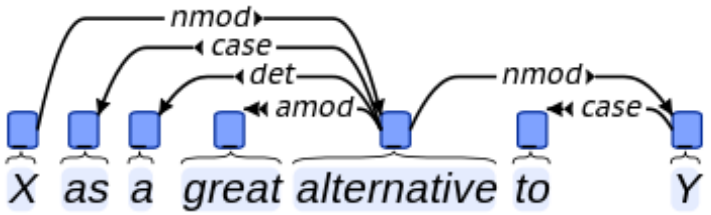
as relating to

as referred to

as commonly referred to

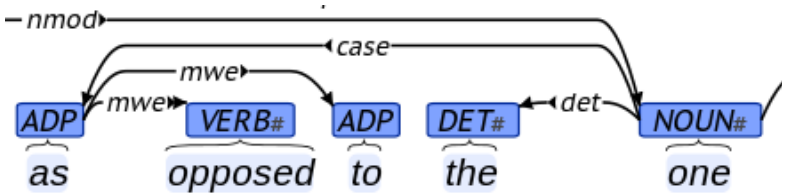
?

- UD: Compositional:



MWE “as opposed to” continuum

- UD: Special relation:



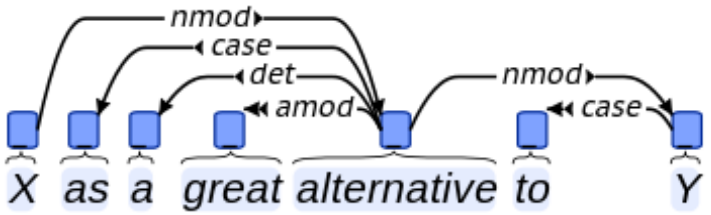
as relating to

as referred to

as commonly referred to

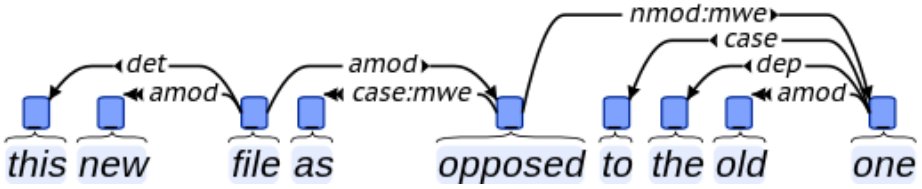
catastrophe

- UD: Compositional:



MWE “as opposed to” complex function names

- standard:special



Labeling
UD:

- *nsubj* and *csubj*
- *nmod* and *amod*

  → –concision, –separability

Labeling UD:

- *nsubjpass:*
  - Dependent is syntactic subject
  - Dependent is noun
  - syntactic subject not first actant of the verb
  - second or third actant
    - A book ←nsubjpass– was given to Craig
    - Craig ←nsubjpass– was given a book

Labeling
get your semantics right

- *it ←subj:0– is raining* (non actancial subject)
- *Ann ←subj:1– gives Craig a book*
- *A book ←subj:2– was given to Craig*
- *Craig ←subj:3– was given a book*

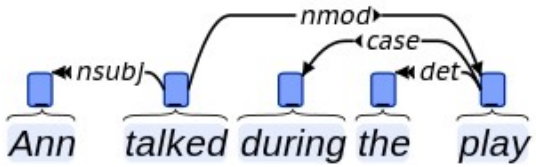
+*separability*

+*transformability*

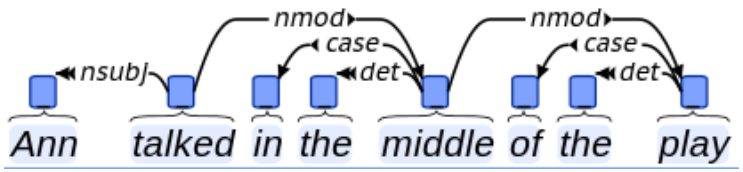
+*level coverage*

Structural choices
Preposition

- Simple preposition

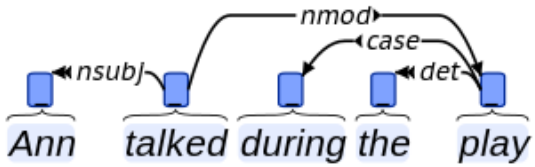


- Complex preposition



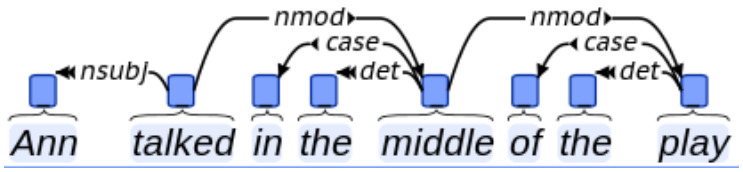
Structural choices
Preposition

- Simple preposition

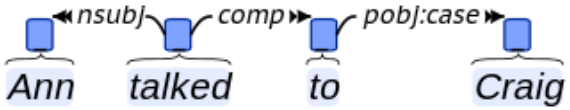


catastrophe

- Complex preposition

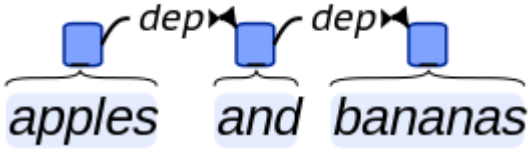


- Proposal:

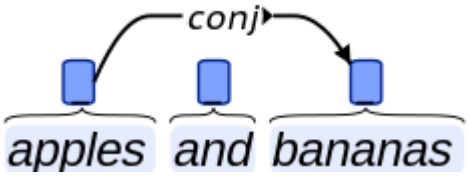


Structural choices
Coordination

- Mel'čukian coordination
  - If conjunction absent: –uniformity

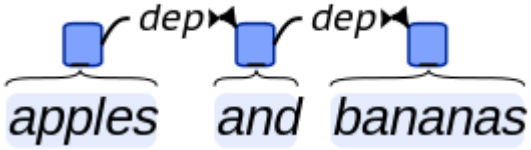


- UD
  - +uniformity
  - adequacy?

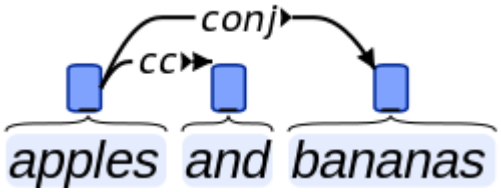


Structural choices
Coordination

- Mel'čukian coordination
  - If conjunction absent: –uniformity

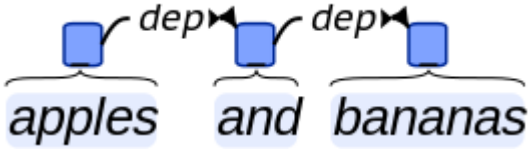


- UD
  - +uniformity
  - adequacy?

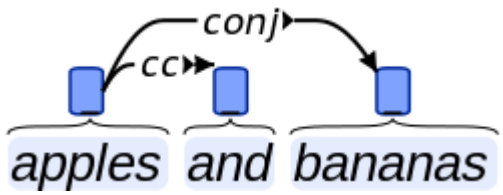


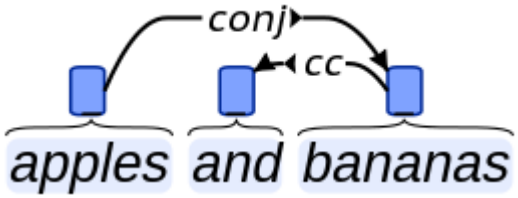
Structural choices
Coordination

- Mel'čukian coordination
  - If conjunction absent: –uniformity



- UD
  - +uniformity
  - –adequacy!
    - [apples and] bananas
    - apples [and bananas]
- Better:





Structural choices Coordination/Disfluency

“I saw a room, a bright room, a room with red lights...”

–conj→

←reparandum–

Structural choices
Coordination/Disfluency

“I saw a room, a bright room, a room with red lights...”

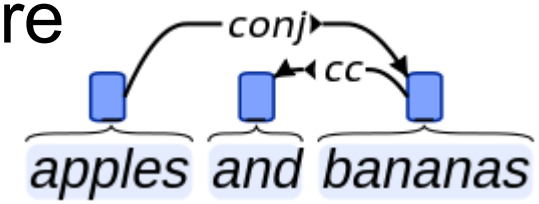
–conj→

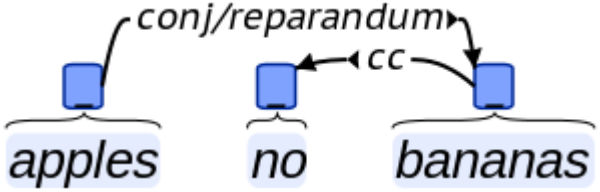
catastrophe

←reparandum–

Structural choices
Coordination

- Coordination in UD is good:
  - Shared dependents:
    - Choice: graph or tree (+*precision*)
  - NCC
    - Choice of tree structure that allows the computation of complete structure
- Improved UD
  - Correct attachment of conjunction
  - Inverse direction of reparandum



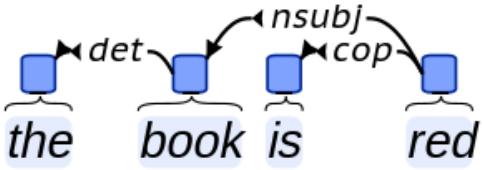


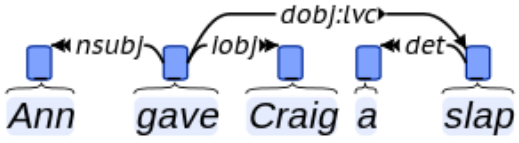
Structural choices
light verbs

- semantically empty words:
  - *a red book* vs. *the book is red*
  - *Ann's slap on Craig* vs. *Ann gave Craig a slap*

Structural choices
light verbs

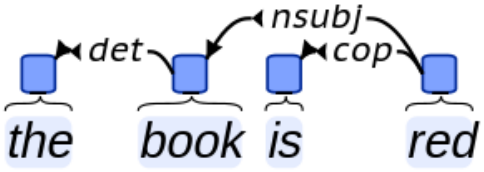
- semantically empty words:
  - *a red book* vs. *the book is red*
  - *Ann's slap on Craig* vs. *Ann gave Craig a slap*
- UD:

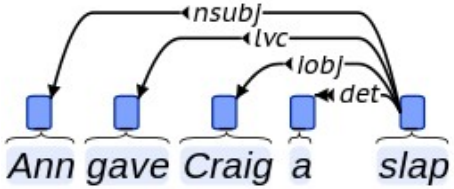




Structural choices
light verbs

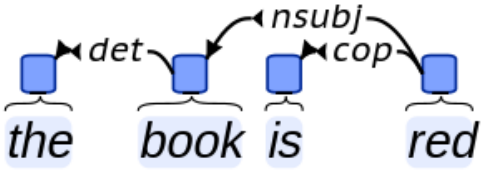
- semantically empty words:
  - *a red book* vs. *the book is red*
  - *Ann's slap on Craig* vs. *Ann gave Craig a slap*
- UD: semantic-centered:

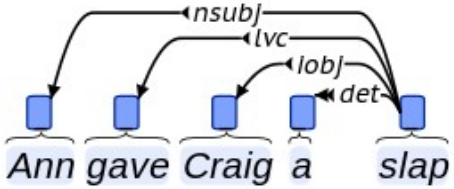




Structural choices
light verbs

- semantically empty words:
  - *a red book* vs. *the book is red*
  - *Ann's slap on Craig* vs. *Ann gave Craig a slap*
- UD: semantic-centered:

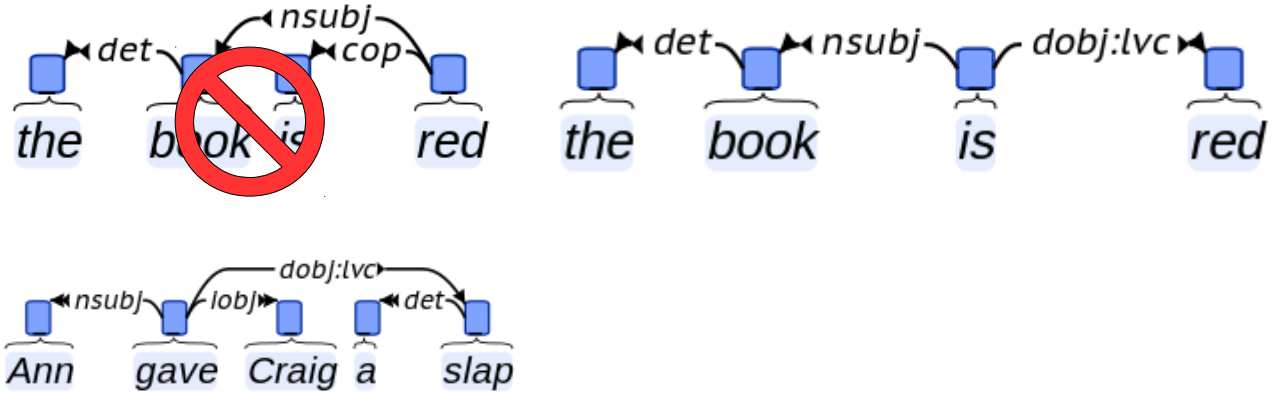




- But: *feel fear* vs. *shake with fear*

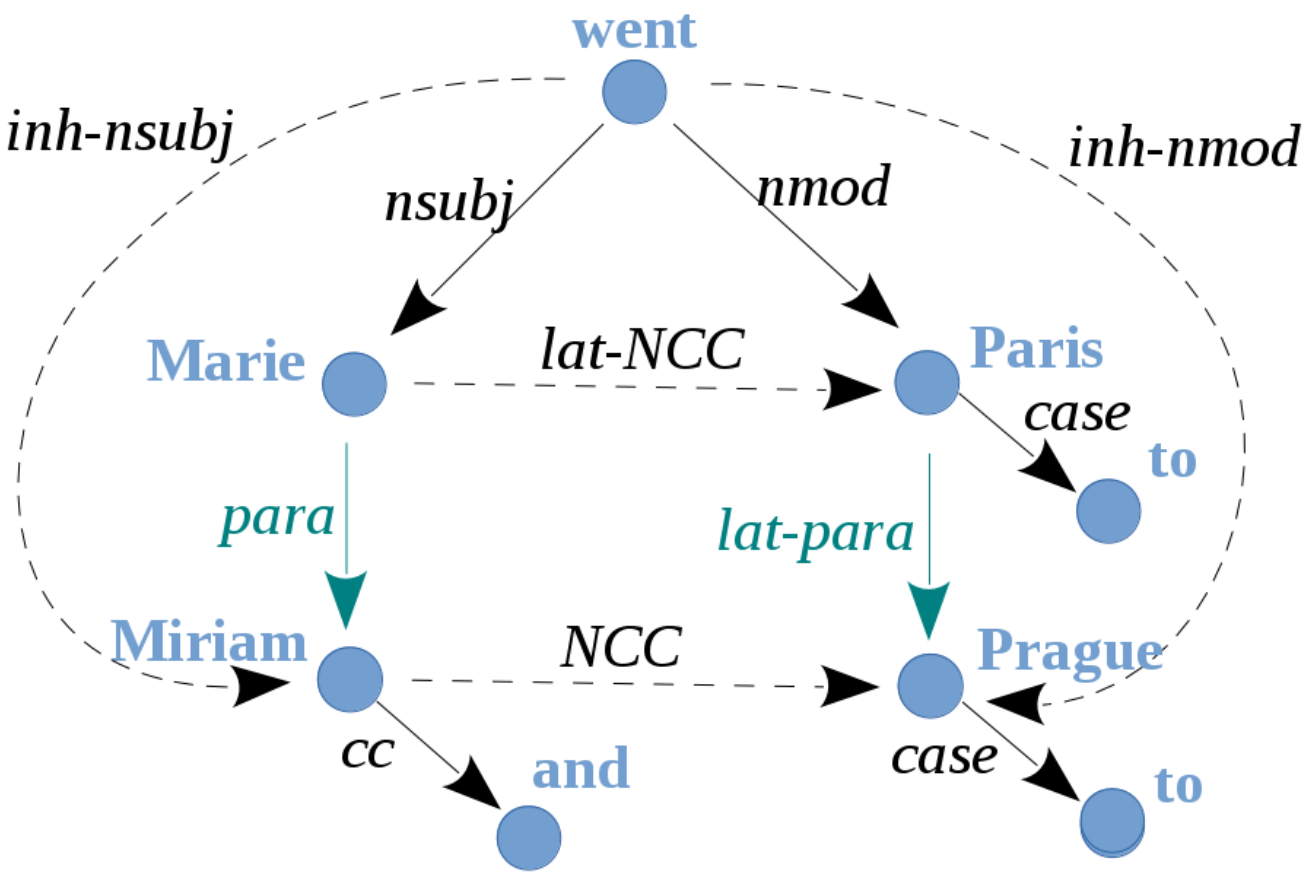
Structural choices
light verbs

- semantically empty words:
  - *a red book* vs. *the book is red*
  - *Ann's slap on Craig* vs. *Ann gave Craig a slap*
- UD: syntax-based:



Non-constituent coordination

NCC structure in “Marie went to Paris and Miriam to Prague” following Gerdes & Kahane (2015), prepositions analyzed in UD style



Comparison with PSG's analyses of coordination

- Symmetric analysis: conjuncts are co-heads (Jackendoff 1977)
  She is [[a linguist]$_{NP}$ and [a computer scientist]$_{NP}$]$_{NP}$

- Asymmetric analysis: the second conjuct is an adjunct (Steedman 1985, Borsley 2005)
  She is [a linguist [and a computer scientist]$_{ConjP}$]$_{NP}$