Speech & Language Processing

Kim Gerdes
Upsay
Limsi

Who and why?

- Who are you?
- Why have you chosen this class? What do you expect?
- What do you know about “Speech and language processing”? What’s your favorite aspect?

more questions to come…

Planning

1 Friday 8/11/24 introduction, syntactic structures Kim Gerdes

2 Friday 15/11/24 Modelling and processing idiomaticity Agata Savary

3 Friday 22/11/24 Expressive Speech Marc Evrard

4 Friday 29/11/24 Machine Learning for Speech Processing: Sequence Discriminative Training Lucas Ondel Yang

5 Friday 6/12/24 General Discussion + From linguistics to NLP Ioana Vasilescu

6 Friday 13/12/24 Paper presentations Kim Gerdes

Next class

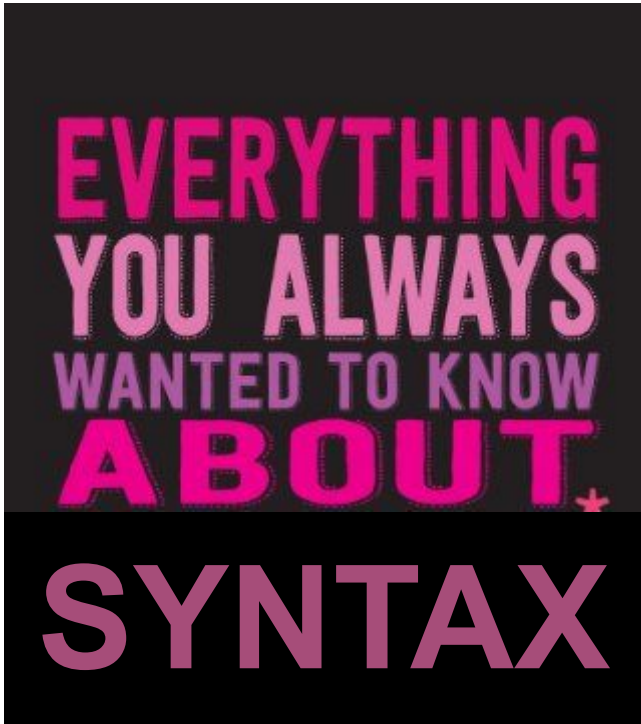
Agata Savary
NLP, semantics, multiwords

Today

- **Those that followed my classes last year, know that I'm very much into patent text generation**
- **But I got a second hobby:**
  - **It starts with S and ends with X**

Today

- **Everything you always wanted to know about syntax but were afraid to ask.**



But why?

- history of AI:
  - idea:
    - understanding the human language faculty will allow us to build better machines
    - what does “understanding” mean?
      - give an example of something that you understand about Language
      - give an example of something that you understand about one particular language

But why?

- history of AI:
  - idea:
    - understanding the human language faculty will allow us to build better machines
    - what does “understanding” mean?
      → rules that humans can grasp
    - Did we collect those rules?
    - Are they discrete/symbolic rules? Why?
    - Did we implement those rules to build AI?

But why?

- History of AI:
  - idea:
    - understanding the human language faculty will allow us to build better machines
  - reality:
    - today’s best AI works by learning what?

But why?

- History of AI:
  - idea:
    - understanding the human language faculty will allow us to build better machines
  - reality:
    - today's best AIs are matrices of billions of parameters
      → not explainable (in legacy/human/rule-based terms)
      → explainable AI? Why?

But why?

- History of AI:
  - idea:
    - understanding the human language faculty will allow us to build better machines
  - reality:
    - today's best AIs are matrices of billions of parameters
    - AI can help us find understandable patterns in the human language faculty

Linguistics → AI ? (maybe later as explainable AI)
AI → Linguistics ! (now)

We still want to understand how Language works.
We have less justification to do so :(
But we have better tools :)


SYNTAX
SYNTAX
SYNTAX
SYNTAX
STANINCE

A few questions

1. Who has been studying grammar?
2. Who has heard of phrase structure?
3. Who has heard of dependency analysis?
4. Who knows what a corpus is?
5. Who thinks something like this when hearing the word “treebank”?
6. Who has heard of typology?

Plan

1) Syntax?
2) Syntactic structures
3) Treebanks
4) Annotation
5) Parsing

1) Syntax?

Linguistics

- Natural or human science?
- linguist's ancestor, the grammarian, was to give access to dead but sacred languages such as Sanskrit and Koranic Arabic
- **Vaugelas (1647):**
  - **His most famous quote: "So here is how we define good Usage. It is the way of speaking of the healthiest part of the Court, in accordance with the way of writing of the healthiest part of the Authors of the time"**

Linguistics

- 19th century:
  - **Colonialism**
  - **Darwinism**
  - **Science and engineering**
- **Humboldt (1836)**
  - **“the lively and inseparable connection between languages and the mental capacity of nations” and has a whole chapter entitled “Less perfect language structures” (§23), starting his description with Semitic languages, Hebrew and Arabic, before moving to Delaware languages.**

“So although we gladly concede that the form of Chinese exhibits, more perhaps than any other language, the power of pure thought, and directs the mind more exclusively and urgently to this, precisely because it lops off all the small distracting sounds of connection, and although the reading of just a few Chinese texts reinforces this conviction to the point of admiration, still, even the most resolute defenders of this language can hardly maintain that it guides the mind’s activity to the true center, from which poetry and philosophy, scientific research and eloquent discourse, spring forth with equal readiness.”

Linguistics

- Take home:
  - As every science: goal driven, serves as justification of power structures
  - Not neutral how to analyze languages
  - Analyzing languages gives them value (Naija)
  - Are all languages equal?
    - Nature vs. Nurture
    - Polite Japanese
    - Geographic Guugu Yimithirr

- **Structural syntax**
  - **Finding simple rules**
  - **Describing the lexicon**
  - **Understanding human language faculty**
  - **Making machines understand human language**

What’s a language

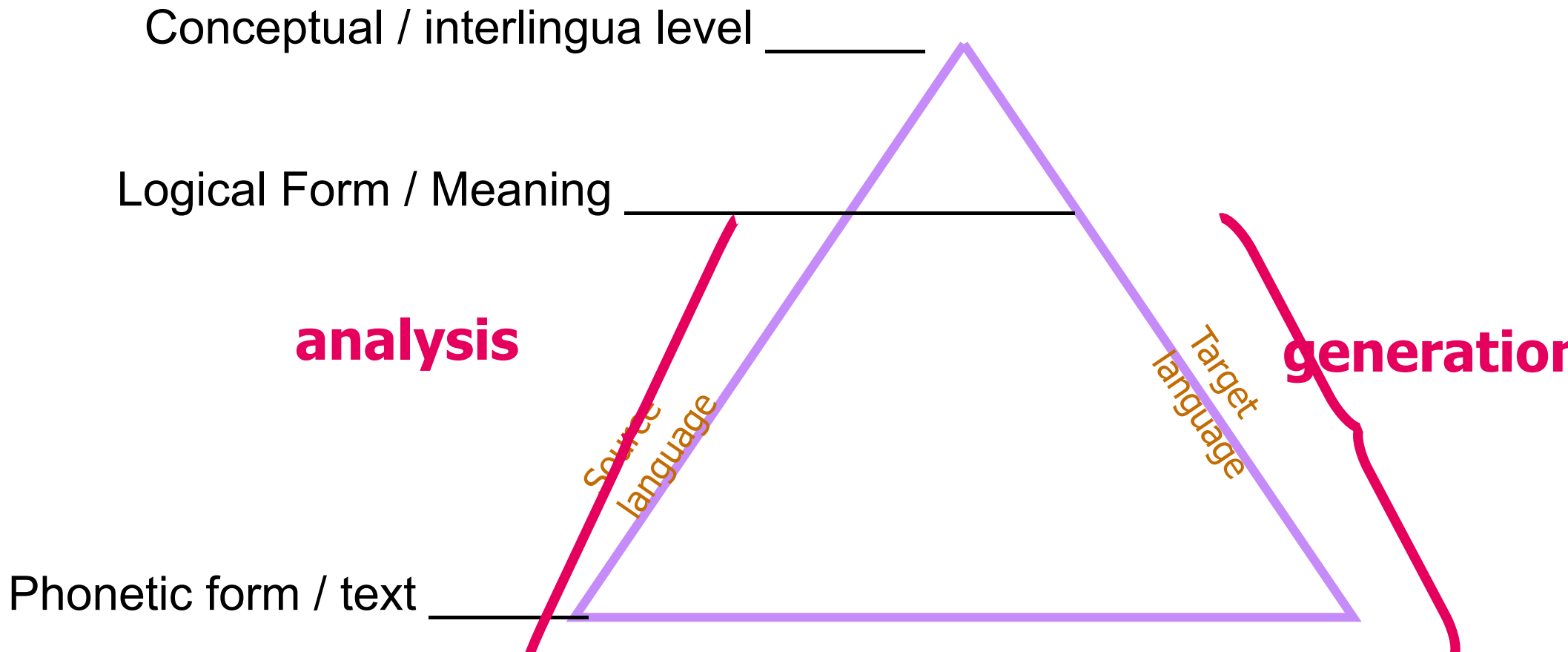
- means of communication between two people (= two brains)

speaker

adressee

- model of a language
- = correspondence between meaning and sound

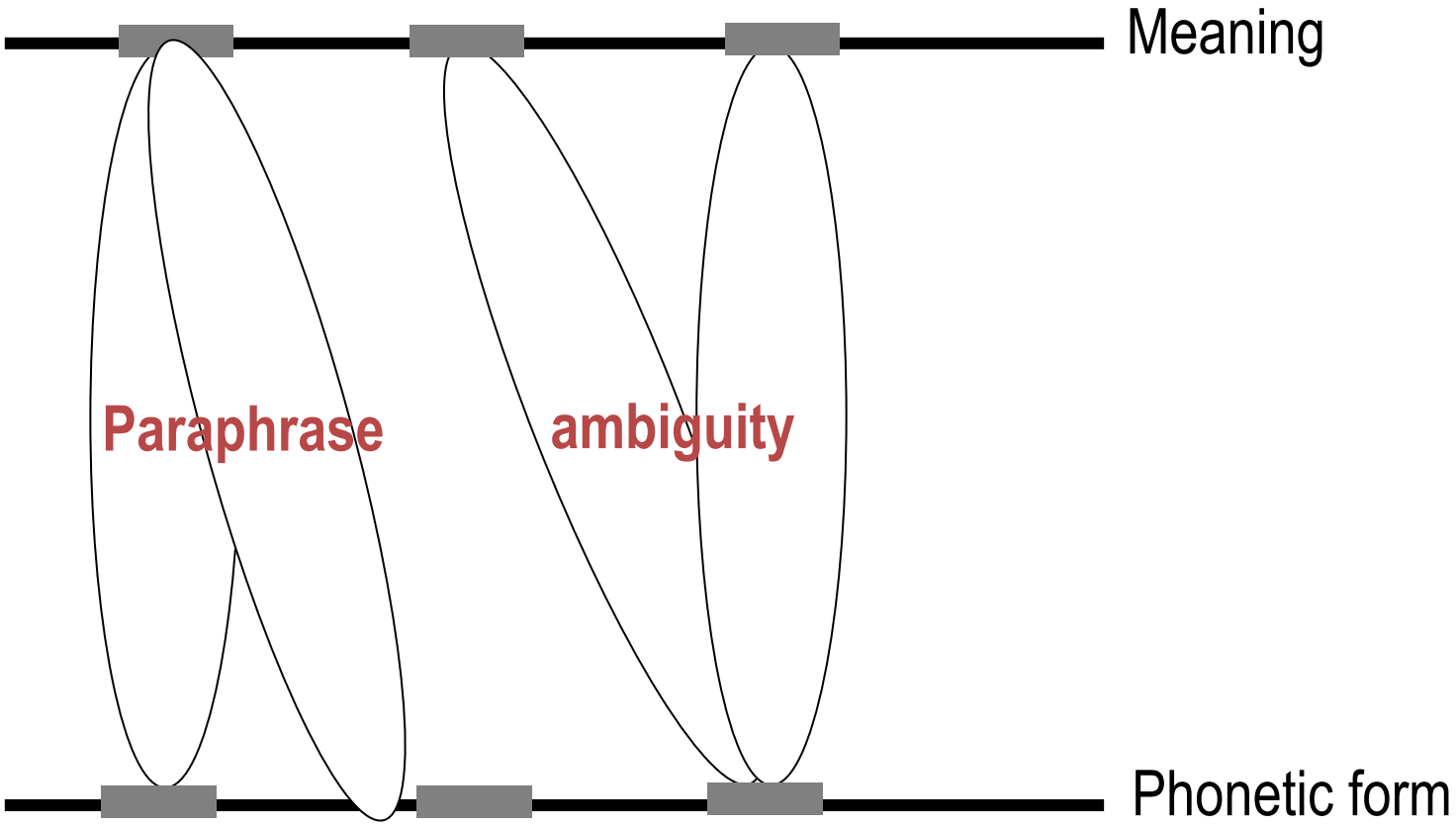
Machine Translation Model
Vauquois' triangle



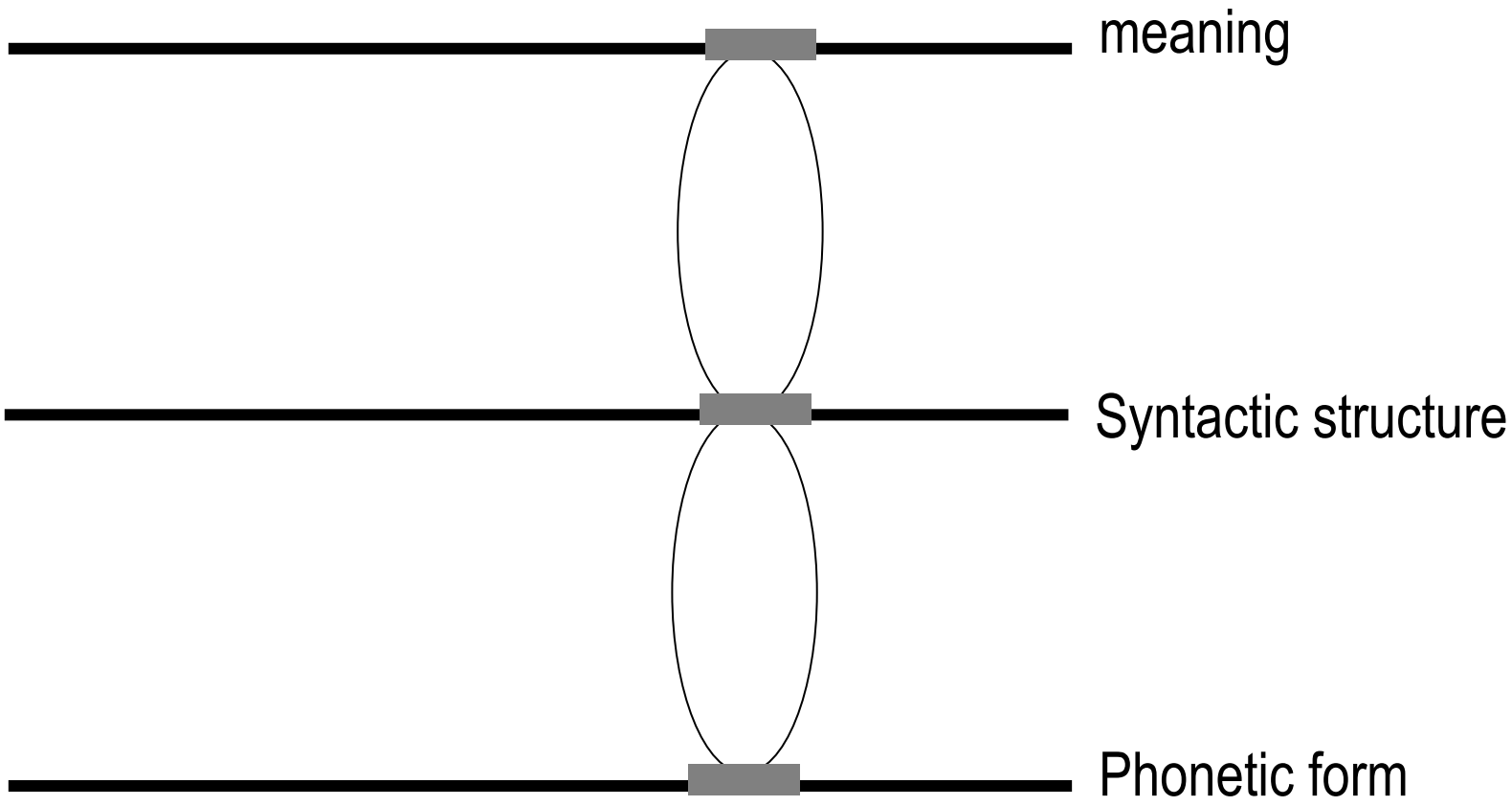
Complete MTT model

- Semantics
- ↕
- Syntax
- ↕
- Topology/Morphology
- ↕
- Prosody

Linguistic model



Syntax



What’s in the syntactic structure

- How the words combine to form a sentence
- Words? Minimal units of syntax
- Sentence? Maximal units of syntax

What is dependency syntax?

- **Syntax, in which the central structure is a dependency tree**

What is a dependency tree?

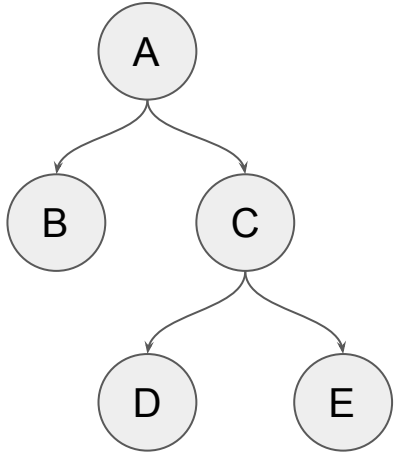
- **A directed graph with functional attributes expressing hierarchical relations between morphemes, words, or semantic units**

- **Put simply:**
  - **Who does what among the words of a sentence?**

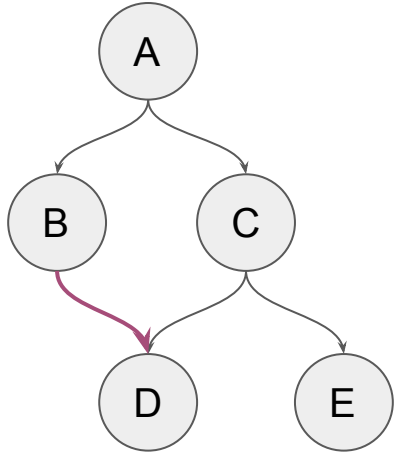
→ A dependency tree is a method to write down which word plays which role in relation to which other word

Can text be folded into trees?

Tree



Graph (DAG)



2) Syntactic structures

What is a treebank?

A **treebank** is a natural language **corpus** annotated:

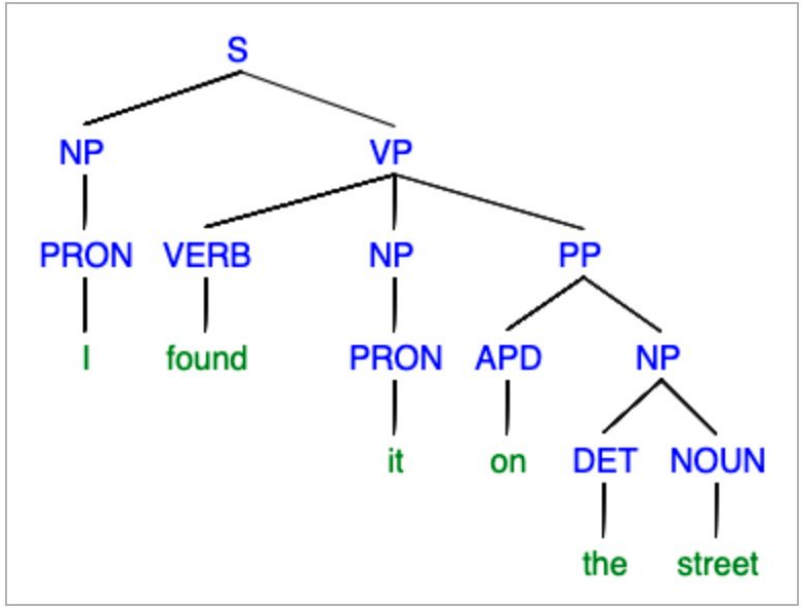
- With a focus on **syntax**
- **Morphology** can also be considered

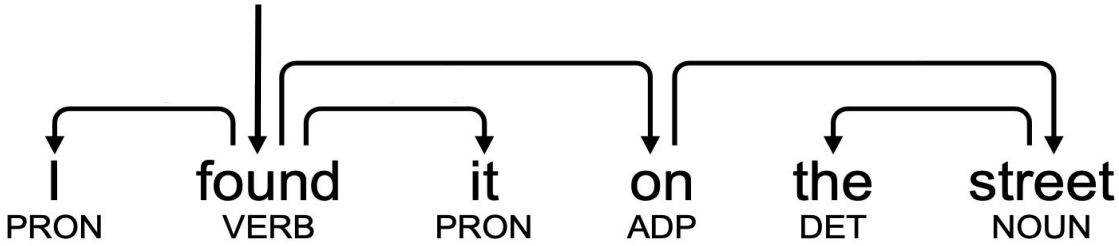
The term “**treebank**”:

- Language data is
  - tokenized (~ **words**)
  - split in smaller unit (called **sentences**)
- The syntax of each sentence is encoded as a **tree**

Syntax: phrase-structures and dependencies

Phrase structure trees: words are recursively grouped together into larger units up to the full sentence

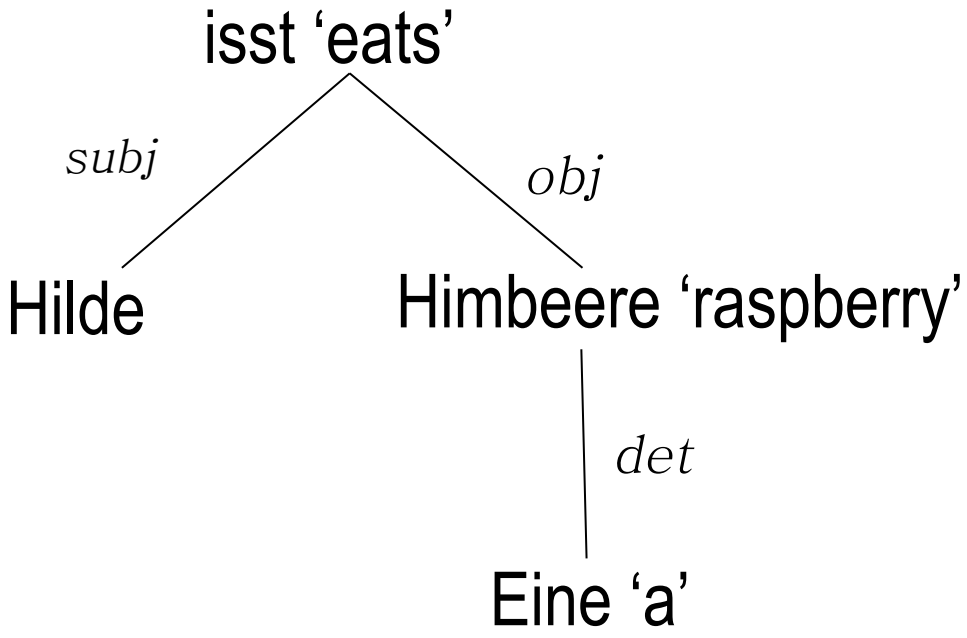




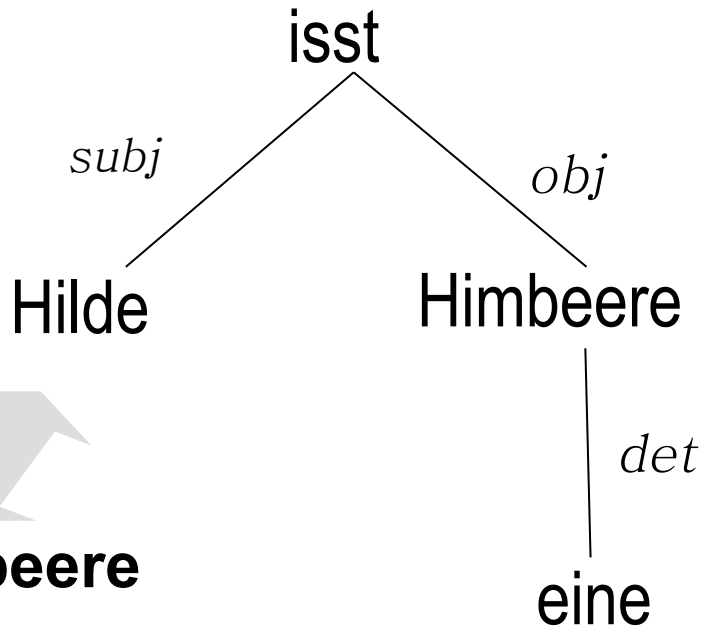
Dependencies: relations are expressed between words. No other units are considered

What is a dependency tree?

- **Example:**

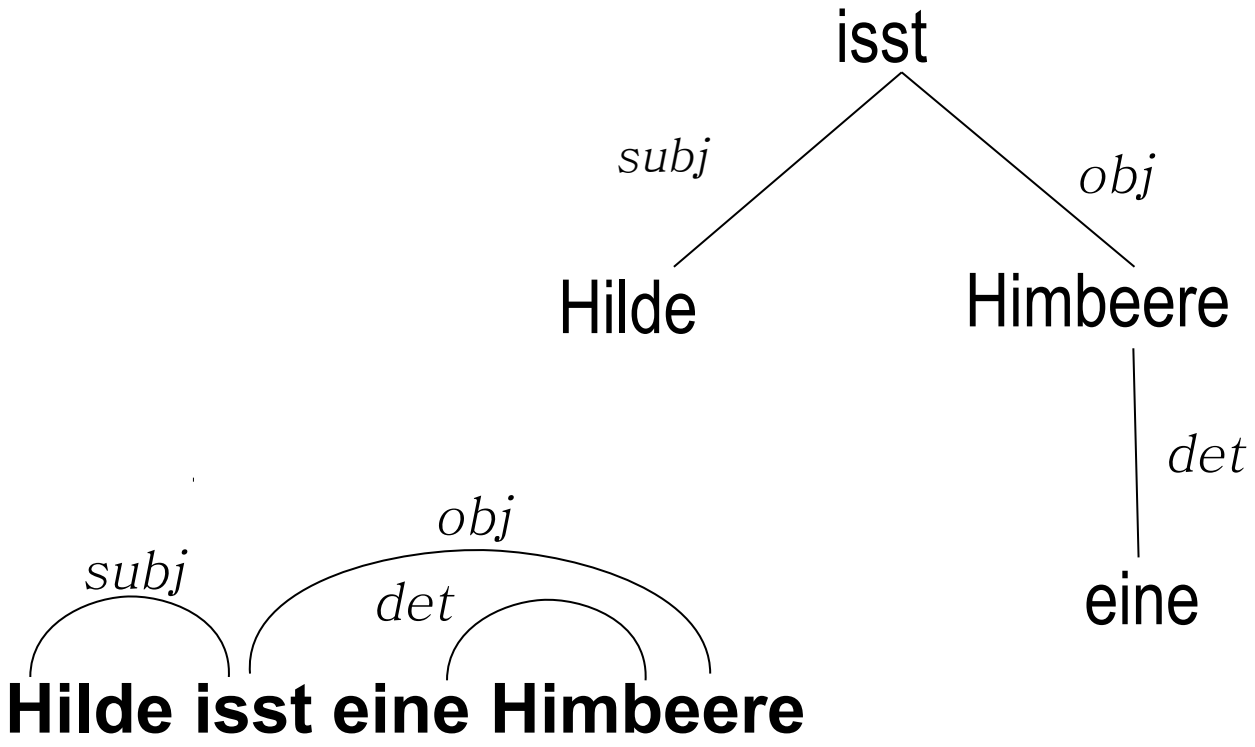


What is a dependency tree?



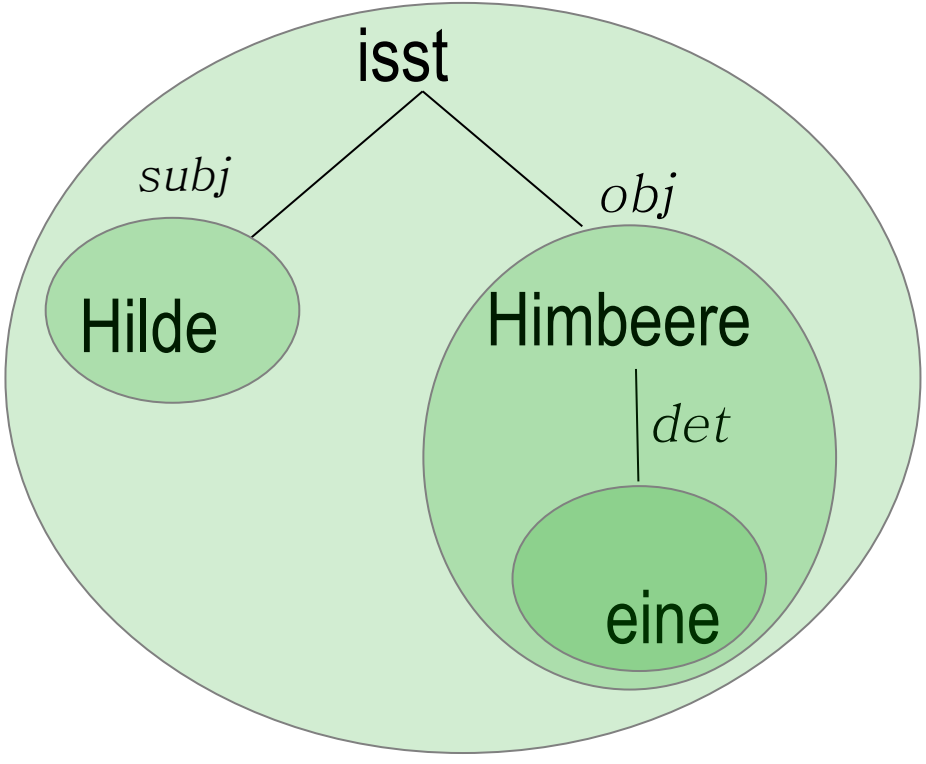
- **Hilde isst eine Himbeere**
- **Eine Himbeere isst Hilde**
- ***'Hilde eats a raspberry / A raspberry is what Hilde eats'***

What is a dependency tree?



What is a dependency tree?

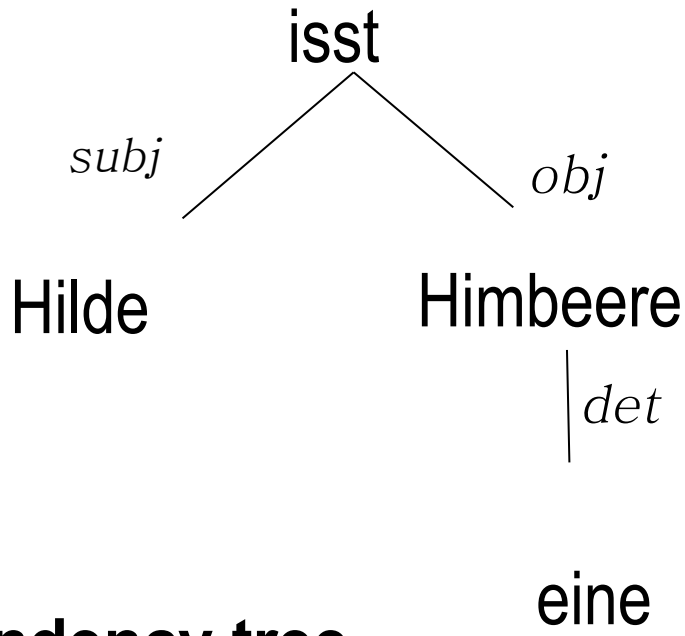
- **Gladkij-tree:**



- **Hypergraph**

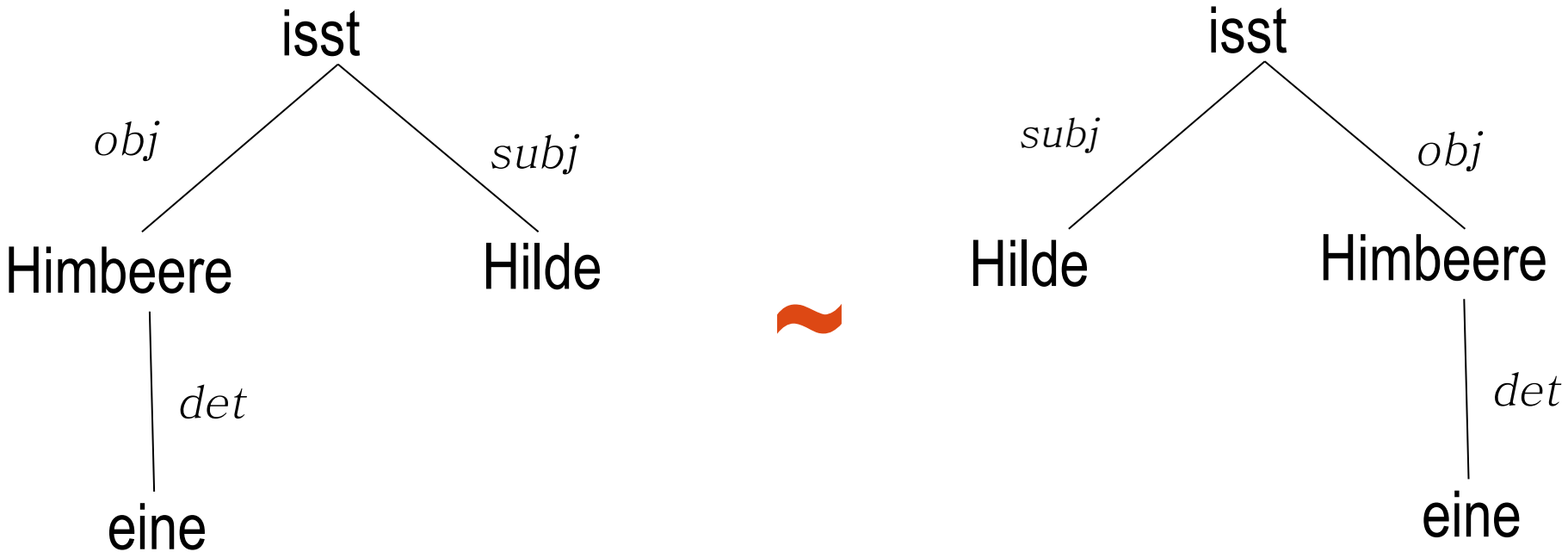
What is a dependency tree?

- **Dependency tree:**

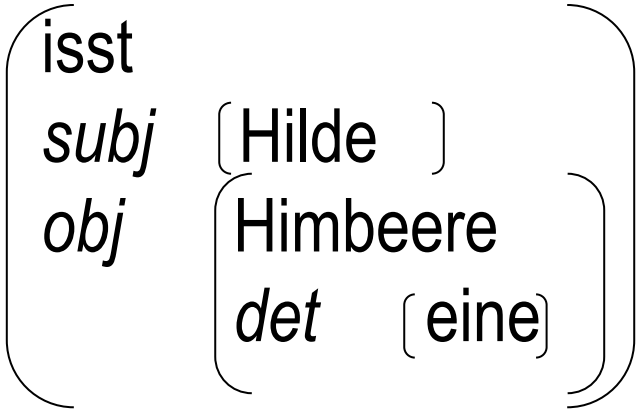


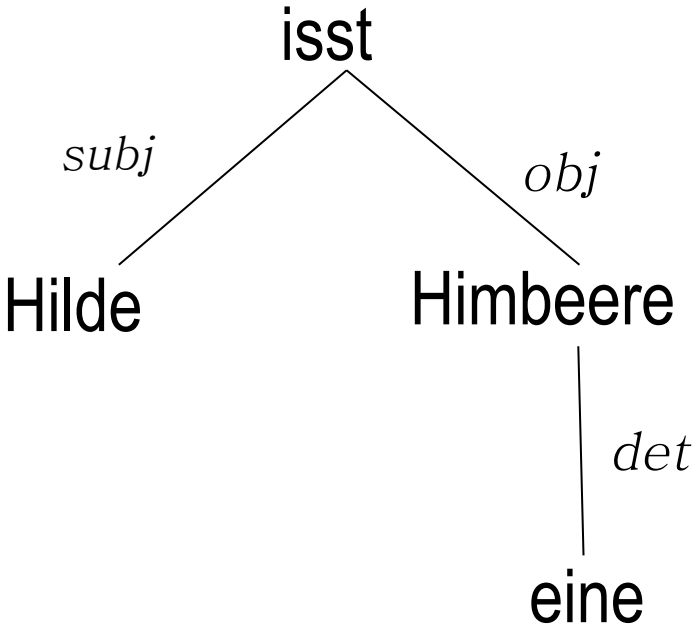
- **Hypergraph ~ Dependency tree**

A dependency tree is unordered



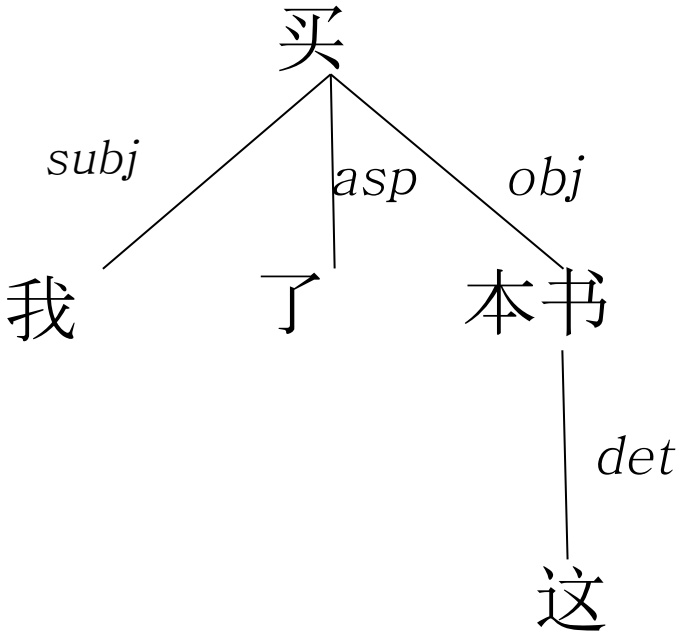
A dependency tree is unordered



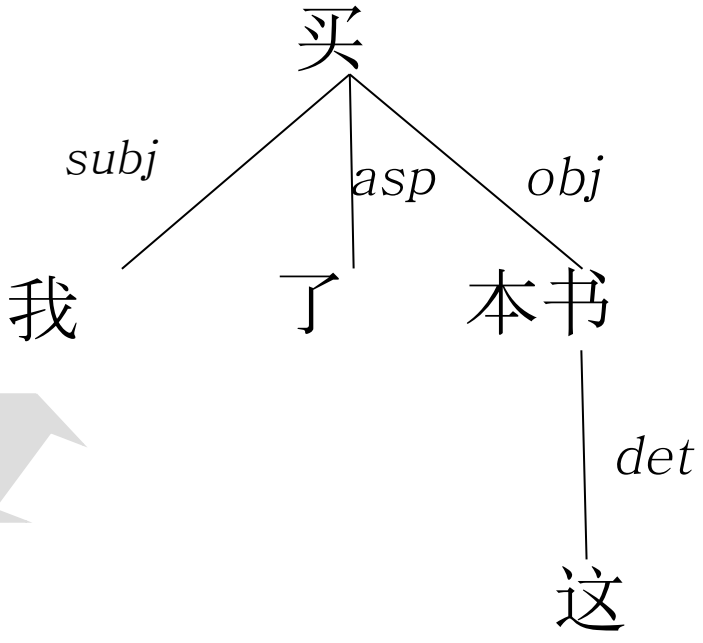


What is a dependency tree?

- **Example:**

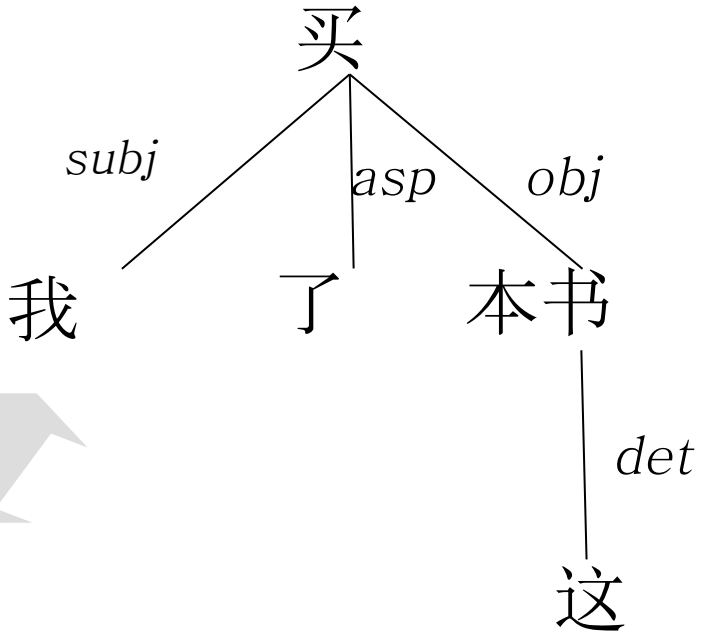


What is a dependency tree?



- **我买了这本书。**
- **Wǒ mǎile zhè běn shū.**

What is a dependency tree?



- 我买了这本书。
- 这本书我买了。

Who does syntax like that?

- **Pāṇini :**
  - **3. century BCE.**
  - **Sanskrit grammar as hierarchical links between words**
- **Ibn Maḍāʾ**
  - **12 century**
  - **Córdoba**
  - **تعلق Ta'alluq = hangs on, depends on, connected with, connection, …**
    - **For the description of relations between verbs and direct oder indirect arguments**

So why is dependency hip?

- **Since about 15 years:**
  - **Dependency is hegemonic in NLP**
  - **Practically no treebanks but dependency treebanks**
  - **Practically no parsers but dependency parsers**

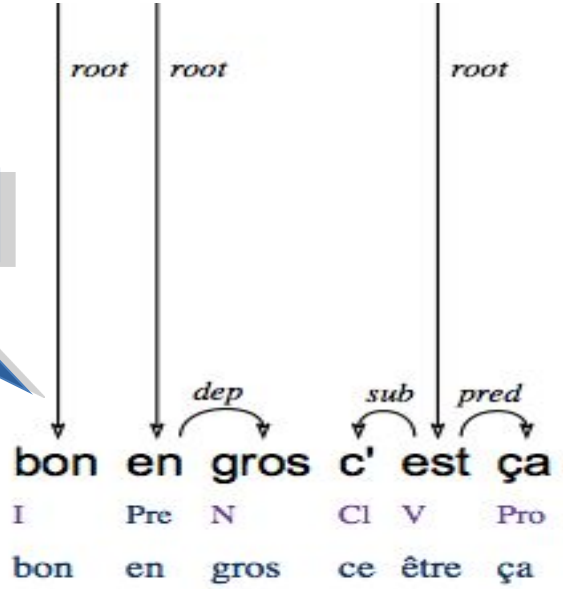
So why is dependency hip?

- 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.

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

Microsyntactic analysis



Dependency and function

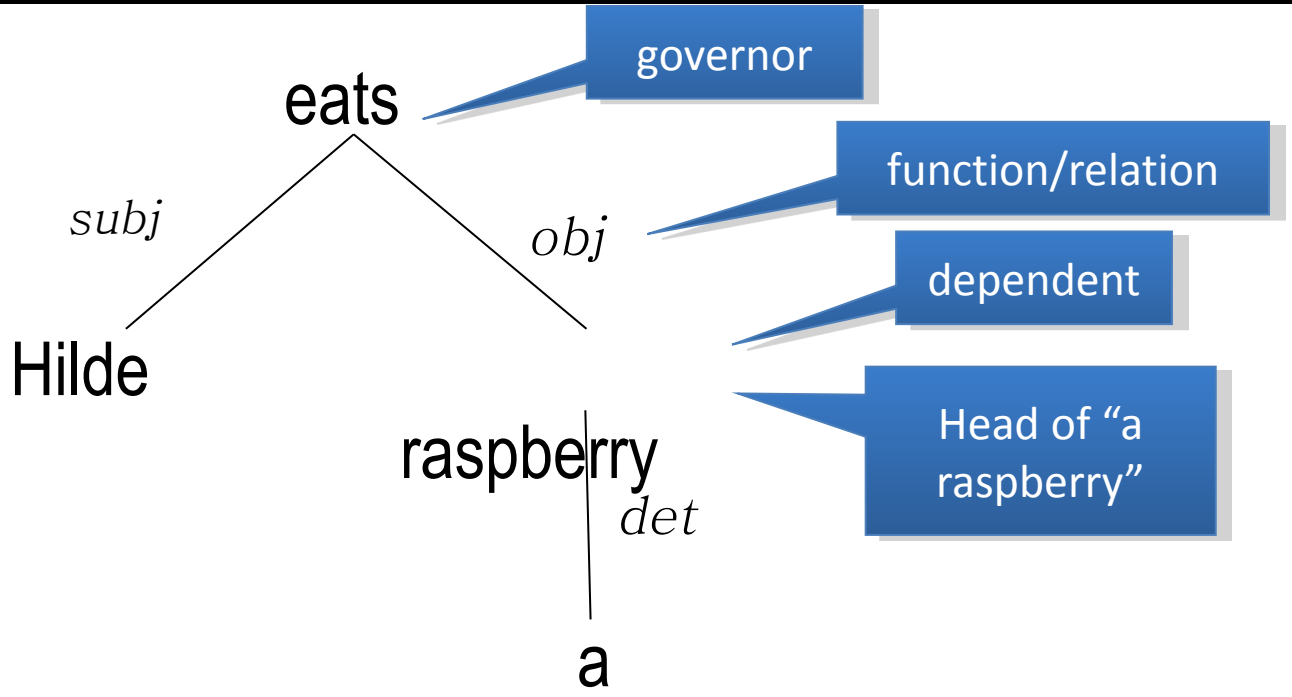
Lexemic segmentation

POS

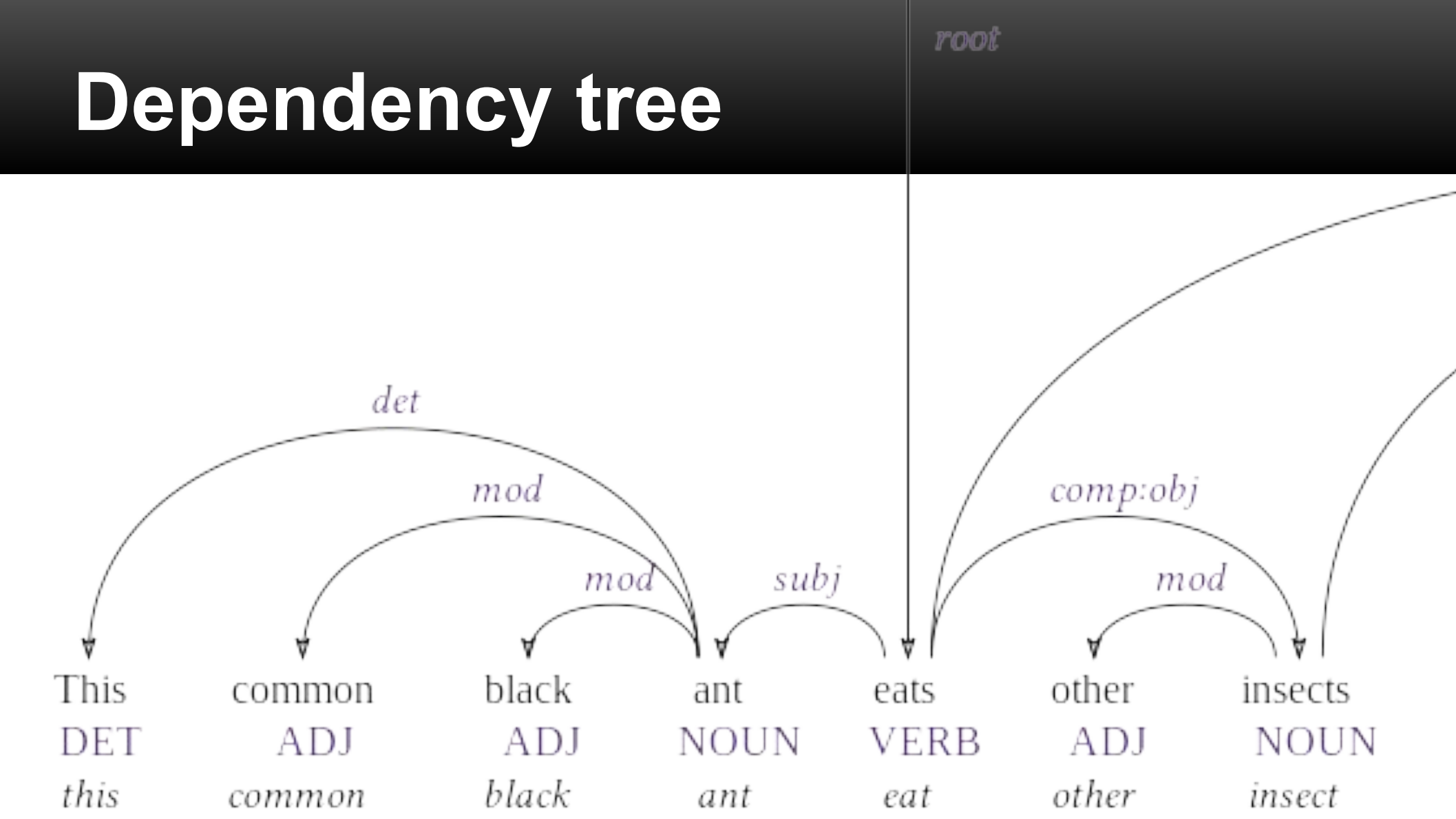
lemmatization

governor dependent head

- **Example:**



Dependency tree

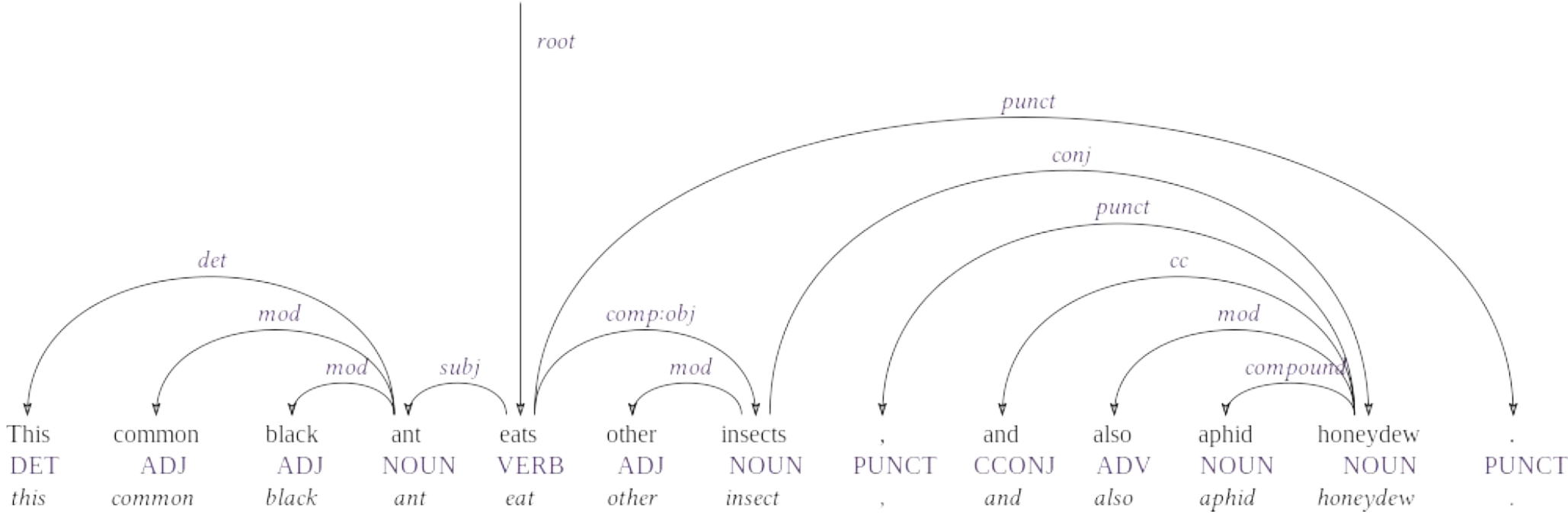


Lemmatization

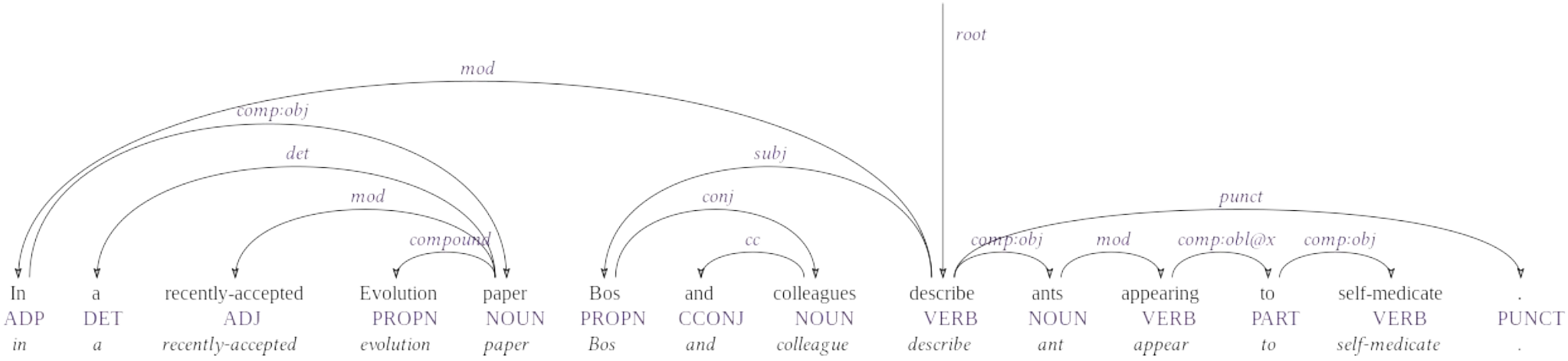
- **The word “to read”:**
  - **English: read, reads, reading (<= 5 forms)**
  - **German: 27 forms (excluding capitalization)**
- **Lemmatization:**
  - **Important for inflectional languages**

gelesen	lesenden
gelesene	lesender
gelesenem	lesendes
gelesenen	lesens
gelesener	lesest
gelesenes	leset
las	lest
lasen	lies
last	liest
lese	läse
lesen	läsen
lesend	läsest
lesende	läset
lesendem	

Dependency tree



Dependency tree



3) Treebanks

Why building syntactic treebanks?

In the pre-digital era:

- For teaching purpose (finding examples of constructions)
- For theoretical purpose (testing a linguistic theory against real examples)

In the pre-LLM era:

- As input and output in NLP tools (training an evaluation of parsers)
- As data for linguistic research

In the LLM era:

- Teaching
- Evaluation of NLP systems
- Data for linguistic research (typology)

Universal Dependencies

https://universaldependencies.org/

UD is an open community effort with over 500 contributors producing over 200 treebanks in over 100 languages.

Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	2	94K	W	IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	<1K		Afro-Asiatic, Cushitic
Belarusian	1	305K	W	IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Bororo	1	<1K		Bororoan
Breton	1	10K	W	IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan
Catalan	1	553K		IE, Romance
Cebuano	1	1K		Austronesian, Central Philippine
Chinese	6	287K	W	Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Chinese	1	433K		Sino-Tibetan
Coptic	1	55K		Afro-Asiatic, Egyptian
Croatian	1	199K	W	IE, Slavic
Czech	5	2,247K	W	IE, Slavic
Danish	1	100K		IE, Germanic
Dutch	2	306K	W	IE, Germanic
English	10	726K	W	IE, Germanic
Erzya	1	20K		Uralic, Mordvin
Estonian	2	528K		Uralic, Finnic
Faroese	2	50K	W	IE, Germanic
Finnish	4	397K	W	Uralic, Finnic
French	7	635K	W	IE, Romance
Frisian Dutch	1	3K		Code switching
Galician	2	164K		IE, Romance
German	4	3,810K	W	IE, Germanic
Gheg	1	15K		IE, Albanian
Gothic	1	55K		IE, Germanic
Greek	2	88K	W	IE, Greek
Guajajara	1	9K		Tupian, Maweti-Guarani
Guarani	1	<1K		Tupian, Maweti-Guarani
Hebrew	2	301K	W	Afro-Asiatic, Semitic
Hindi	2	375K	W	IE, Indic
Hittite	1	1K		IE, Anatolian

Universal Dependencies

- universaldependencies.org
- 259 treebanks in 148 languages (November 2023)
  - Very different size and quality
  - Same annotation scheme
  - based on Stanford dependencies, Google universal part-of-speech tags, Interset interlingua
  - Attempts not to be Anglo/Euro-centric
  - Mostly Indo-European languages, agglutinating languages, few African and American languages.
- Since 2013
- i’m involved in: Spoken French, spoken Mandarin, legal Mandarin, spoken Cantonese, spoken Naija, Old French

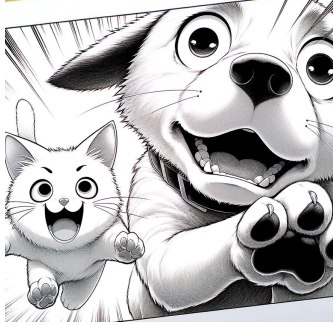
Universal to do what?

- 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

SUD: Surface-syntactic Universal Dependencies

- new surface-syntactic **annotation scheme** similar to UD
- links and dependency labels defined based on purely syntactic **distributional criteria,** giving dependency structures closer to traditional dependency syntax
  - *Meaning-Text Theory, Mel'čuk 1988; Word Grammar, Hudson 1984, 2007; Prague Dependency Treebank, Hajič et al. 2017*
- elementary **conversion** going both ways
  - Goal: without loss, i.e. an "isomorphic" annotation
  - Reality: near-isomorphic http://grew.fr: a Graph Rewriting Tool (Guillaume et al. 2012)
  - Grammars UD to SUD and SUD to UD
- References:
  - UD workshop, Gerdes et al. 2018, Gerdes et al. 2019
  - https://surfacesyntacticud.github.io/
  - Naija guidelines: https://surfacesyntacticud.github.io/guidelines/pcm/

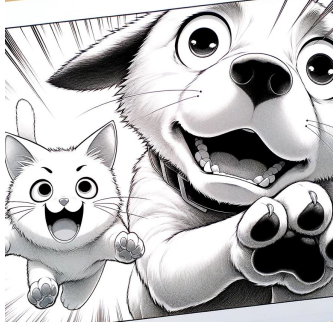
Surface Syntactic UD example

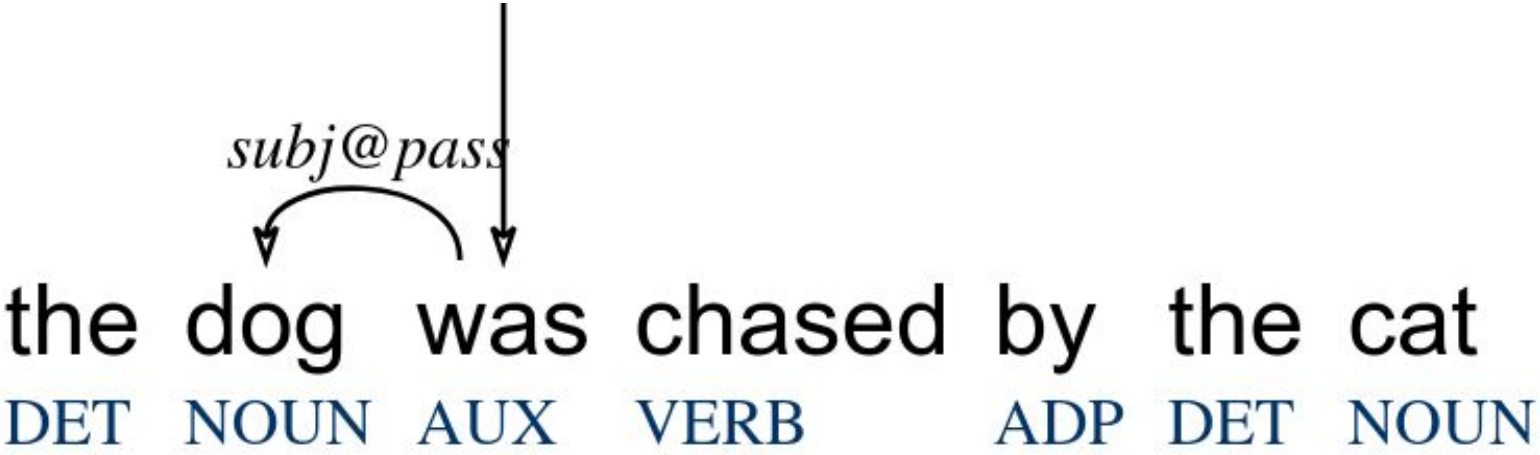


the dog was chased by the cat

- Subject?
- Of what?
- Agreement?
- Name of construction?

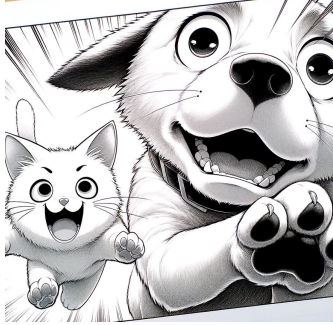
Surface Syntactic UD example

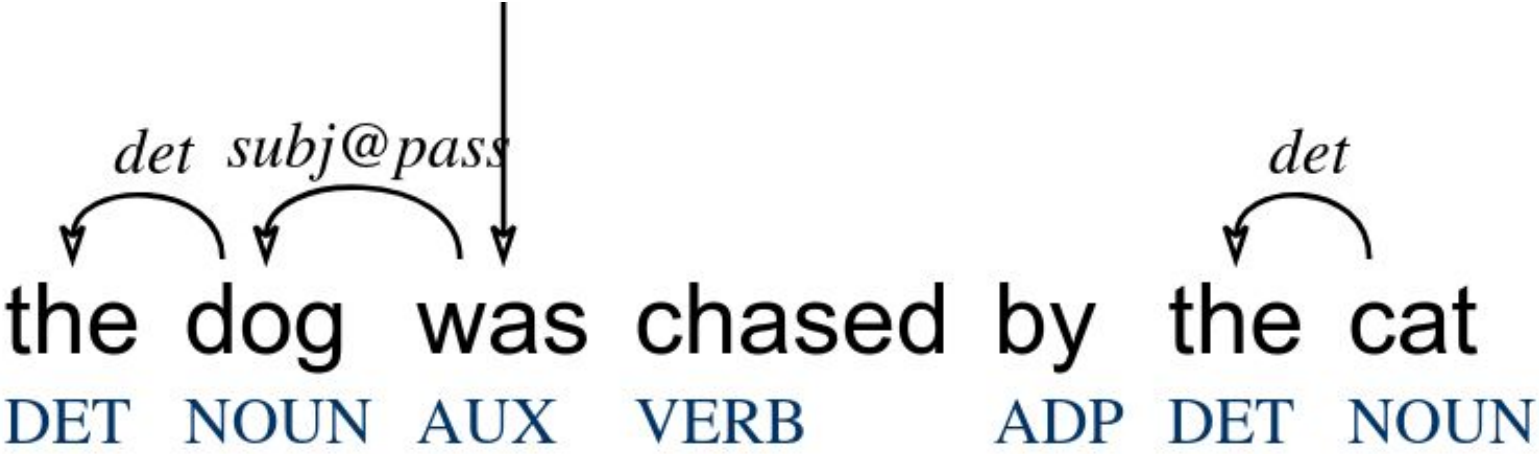




- What else is easy?
- Determiners!

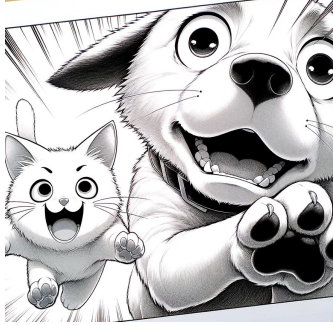
Surface Syntactic UD example

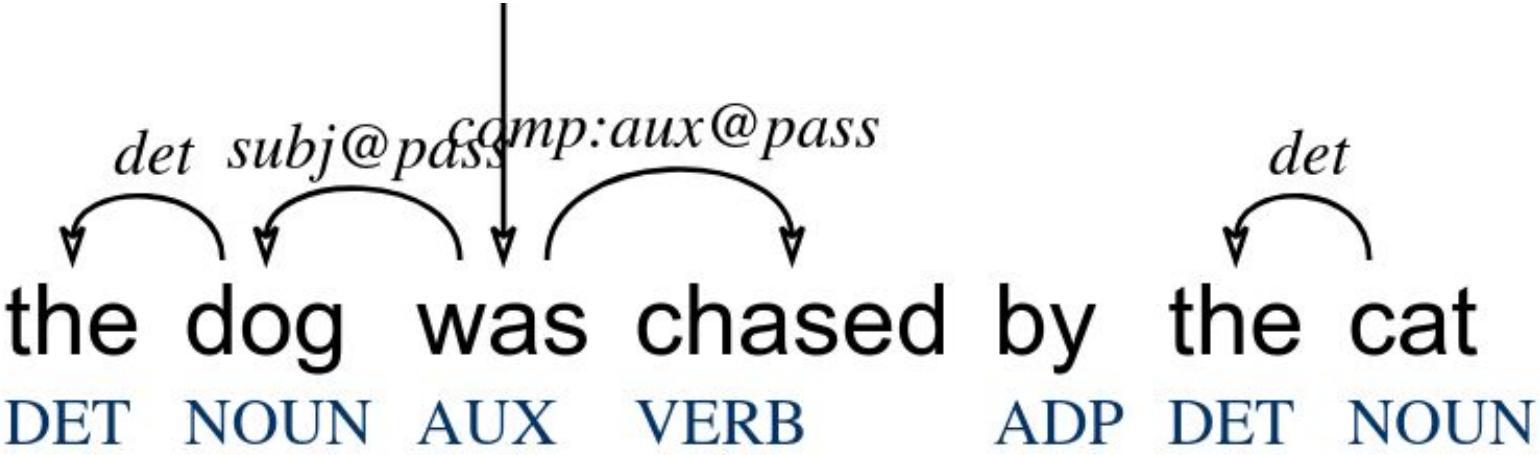




- “was chased” forms a phrase
- Which relation?

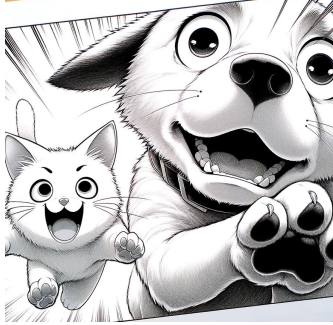
Surface Syntactic UD example

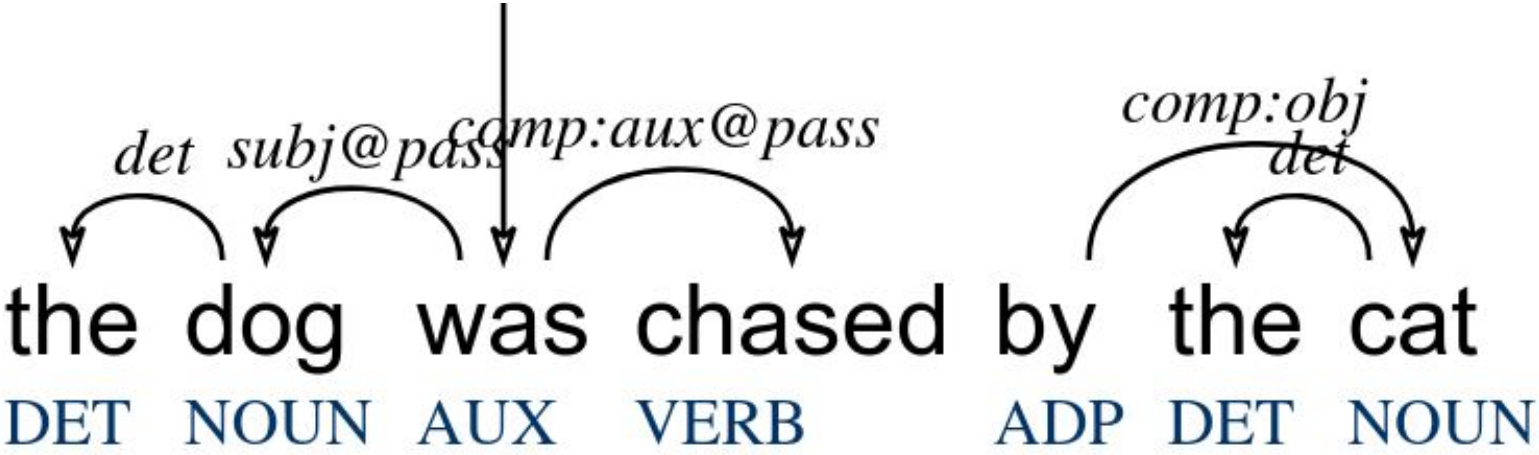




- Complex relation: **comp:aux@pass**
  - complement, auxiliary relation, passive construction
- What kind of phrase is “by the cat”?

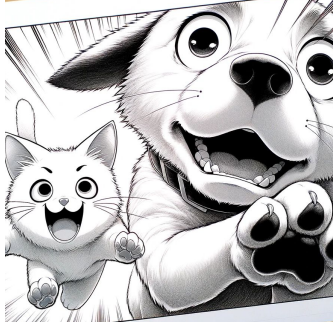
Surface Syntactic UD example

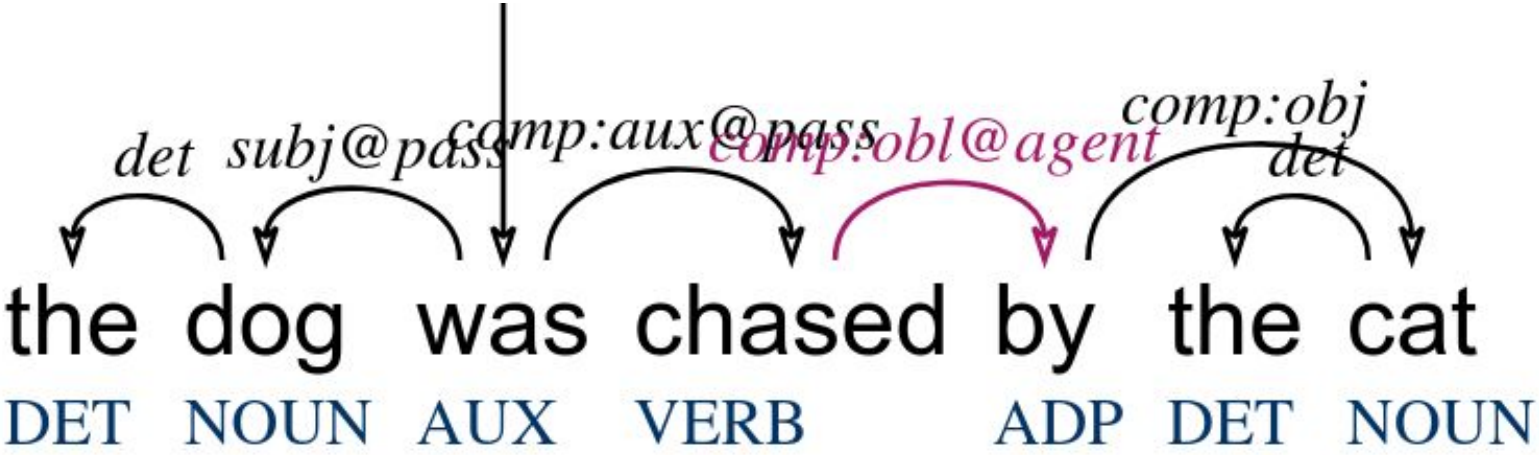




- “by the cat” is a Prepositional Phrase (PP)
- What does “by the cat” do in this sentence?

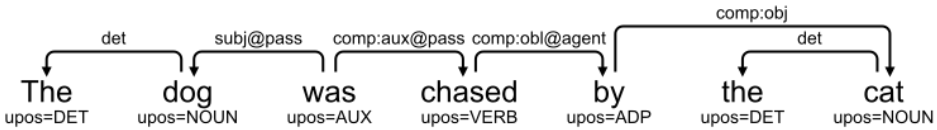
Surface Syntactic UD example

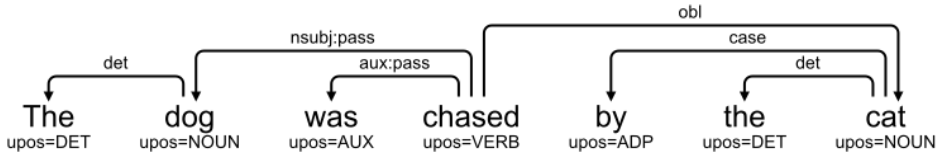


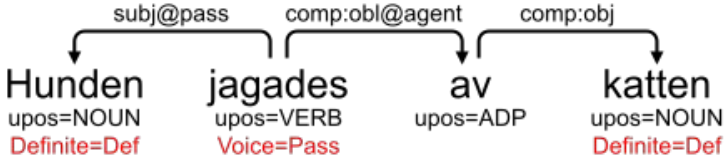


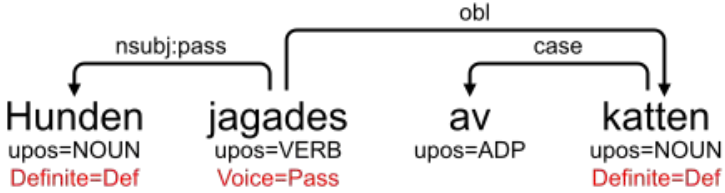
- What does “by the cat” do in this sentence?
  - It’s an oblique complement
  - Semantically it contains the agent
  - → **comp:obl@agent**

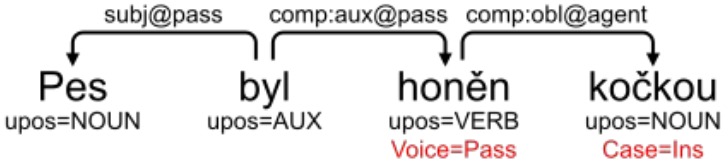
Surface Syntactic UD

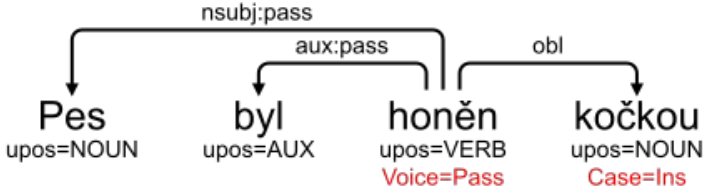




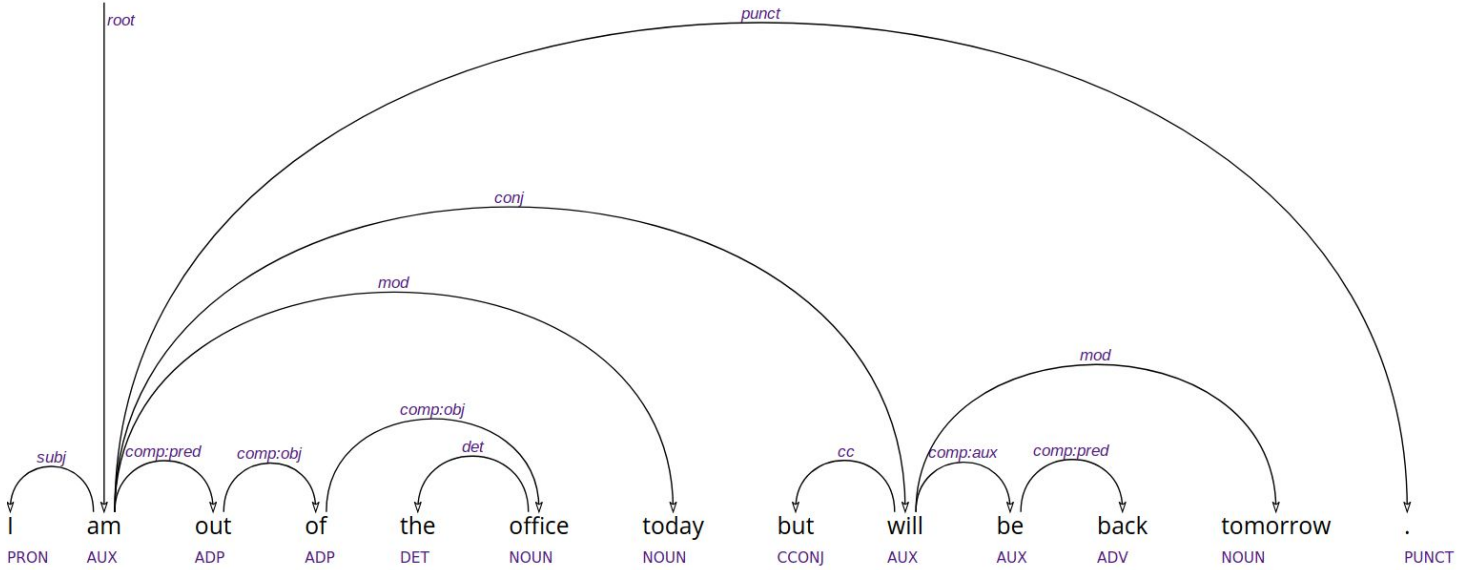




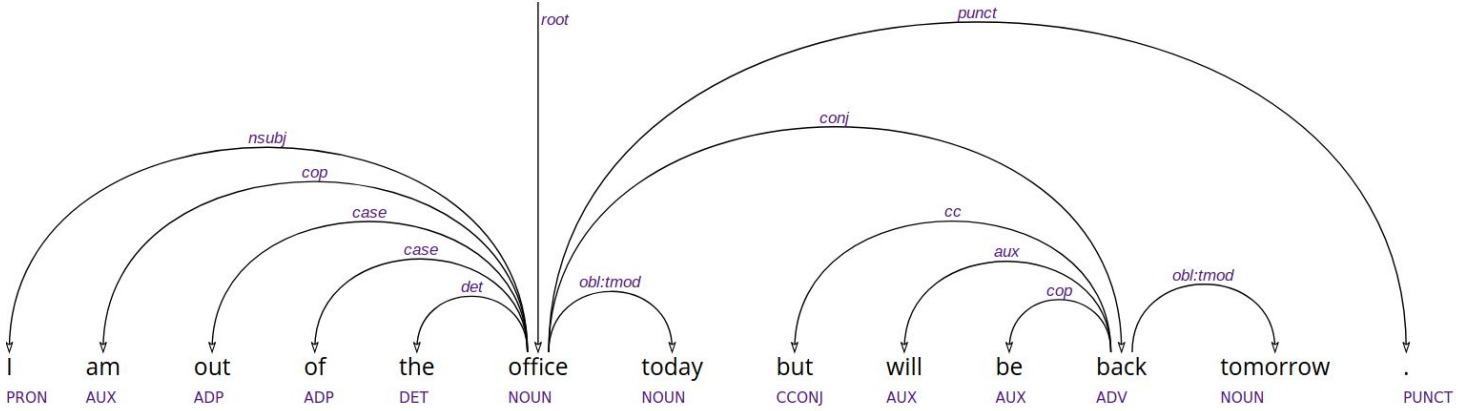




I am out of the office today but will be back tomorrow .

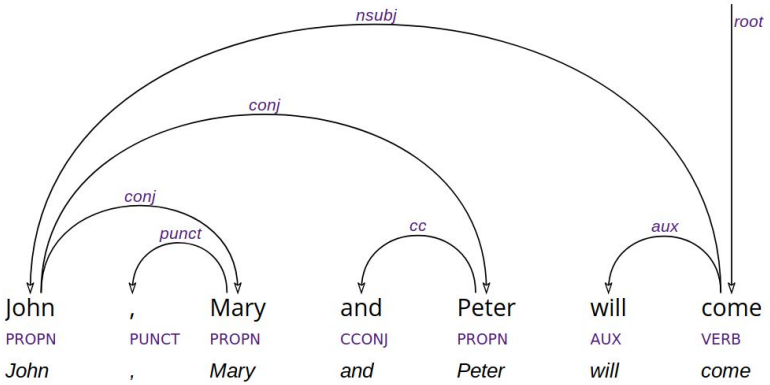


I am out of the office today but will be back tomorrow .

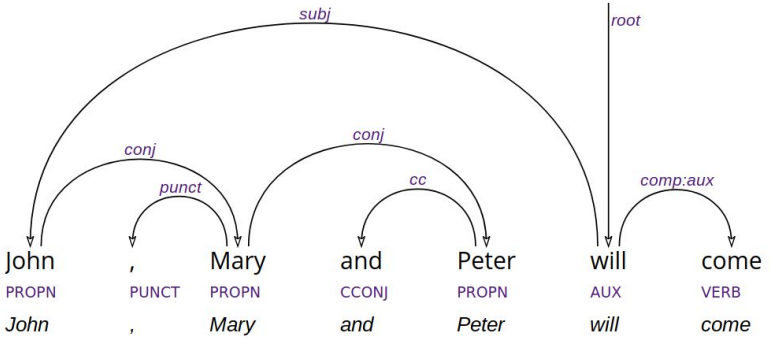


SUD: Differences: Deeper trees: coordination

John , Mary and Peter will come



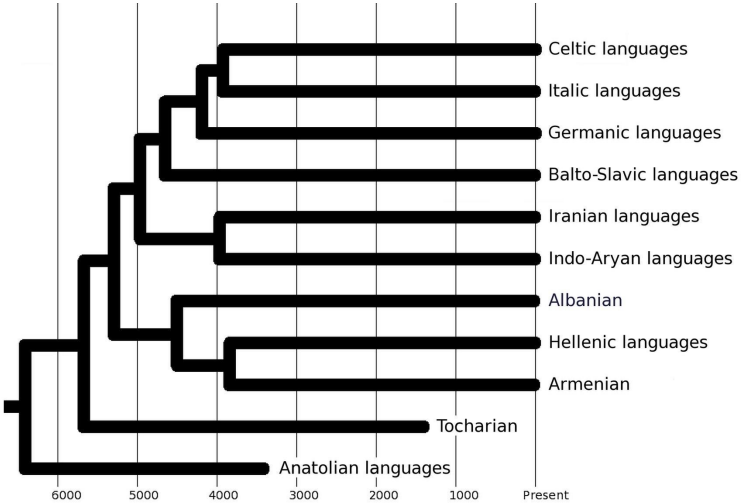
John , Mary and Peter will come

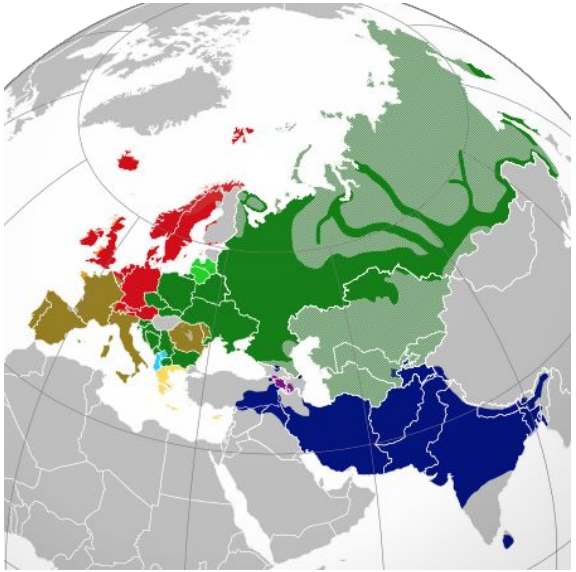


Typometrics

Treebanks to do typology

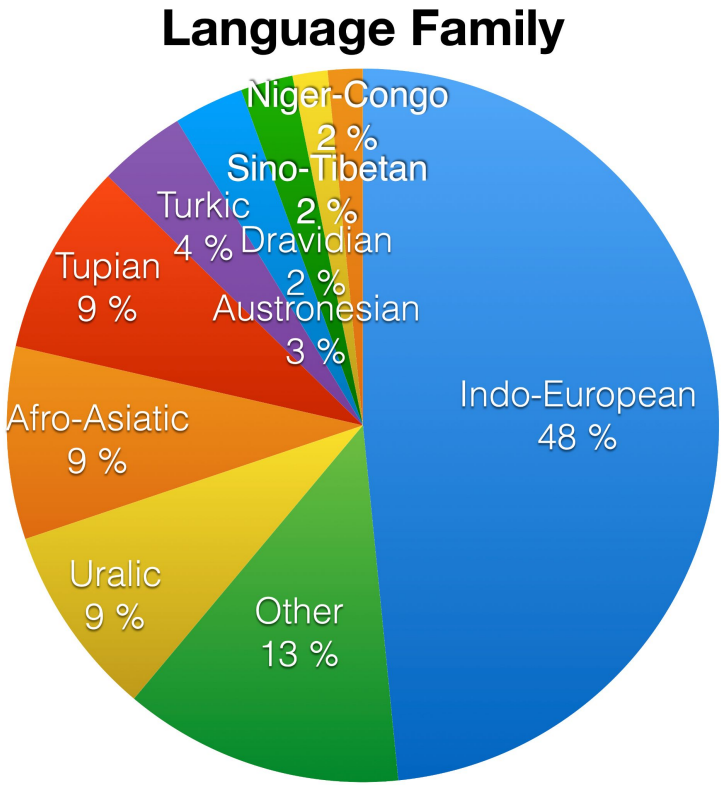
- **Typology: study of how language structure differs**
- **Old linguistic discipline (Indo-European languages, Humboldt)**

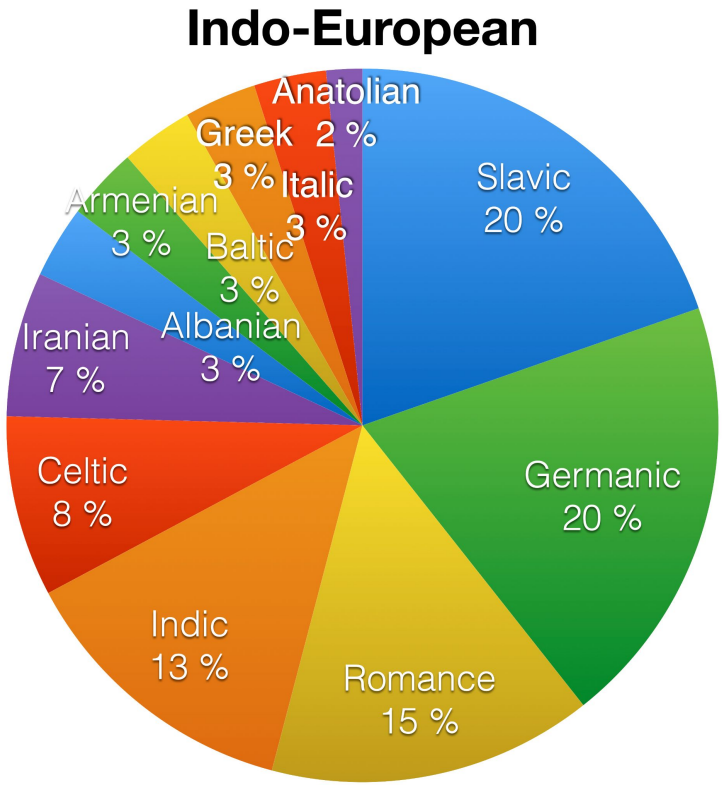




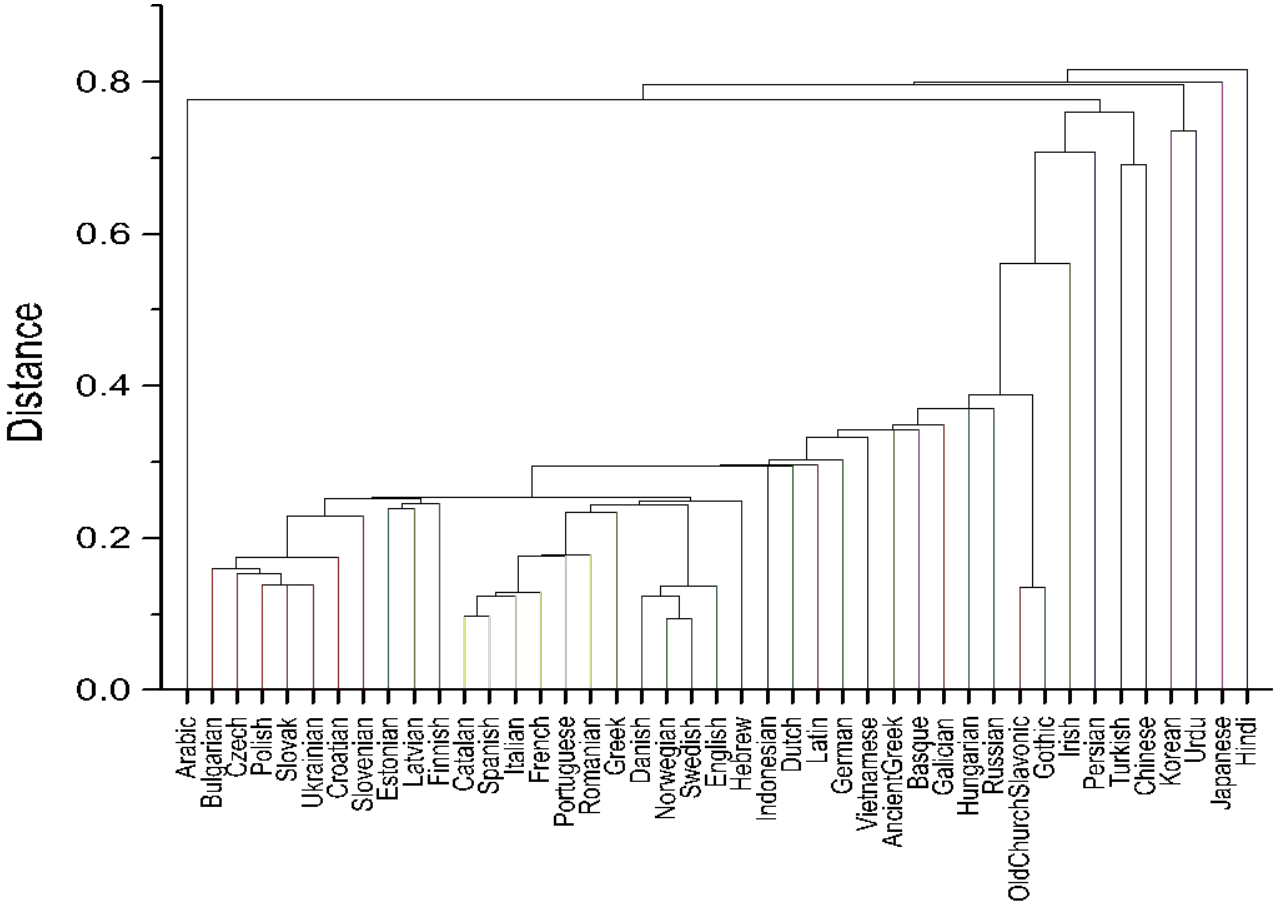
Languages in UD data

Source: “Tutorial on Universal Dependencies” (de Marneffe, Nivre, Zeeman). https://github.com/UniDive/2023-unidive-webinar



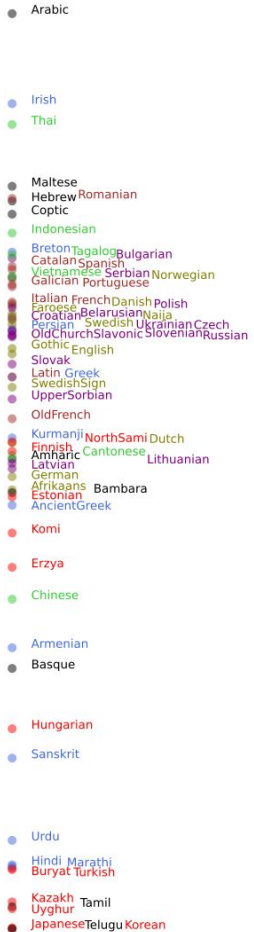


Dendrogram of distance × relative frequency: per language



Typometrics: Tesnière's 1959 "typological classification of languages according to the nature of linearization"

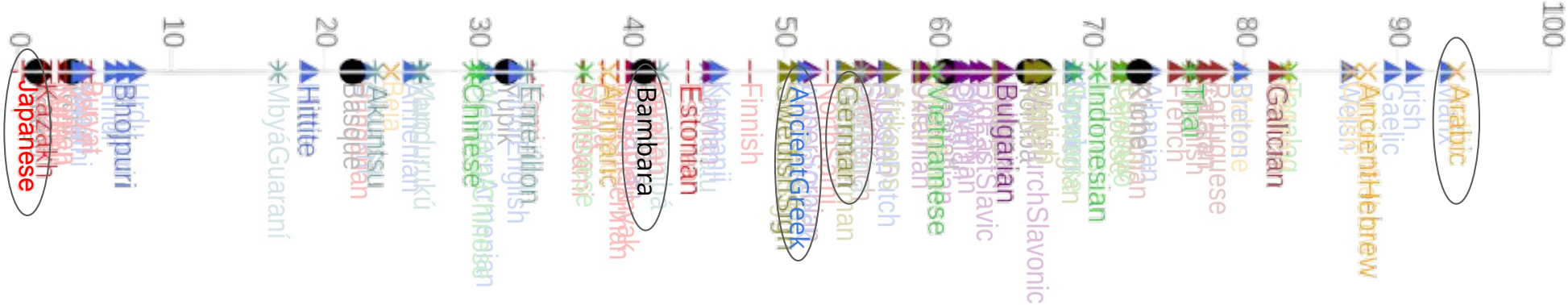
- **Languages**
  - **centrifugal**
    - strict
      - Bantu languages
      - Semitic languages (Hebrew, Arabic)
      - Austronesian languages (Malaysian, Polynesian)
    - mitigated
      - Black-African languages
      - Tasmanian
      - Old Asian languages (Elamite, Sumerian)
      - American languages
      - Basque
      - Indo-European languages
        - Celtic languages
        - Romance languages (Italian, Spanish, **French**)
  - **centripetal**
    - mitigated
      - Indo-European languages
        - Greek
        - Latin
        - Germanic languages (English, **German**)
        - Slavic languages (Russian)
      - Languages of Australia
      - Papuan languages
      - Languages of Andaman islands
      - Sino-Tibetan languages (Chinese)
    - strict
      - Caucasian languages
      - South-African languages (Khoekhoe, Bushman)
      - Dravidian languages
      - Burushaski
      - Hyperborean languages
      - Ural-Altaic languages (including Japanese and Korean)



Head-initiality(L)

head_initiality(L, C) =
% of occurrences of C in L that are head-initial
(governor < dependent)

head_initiality(L)
= **weighted average** of *head_initiality(L, C)* on C



Weighted Average Head-initiality of core relations across SUD 2.11

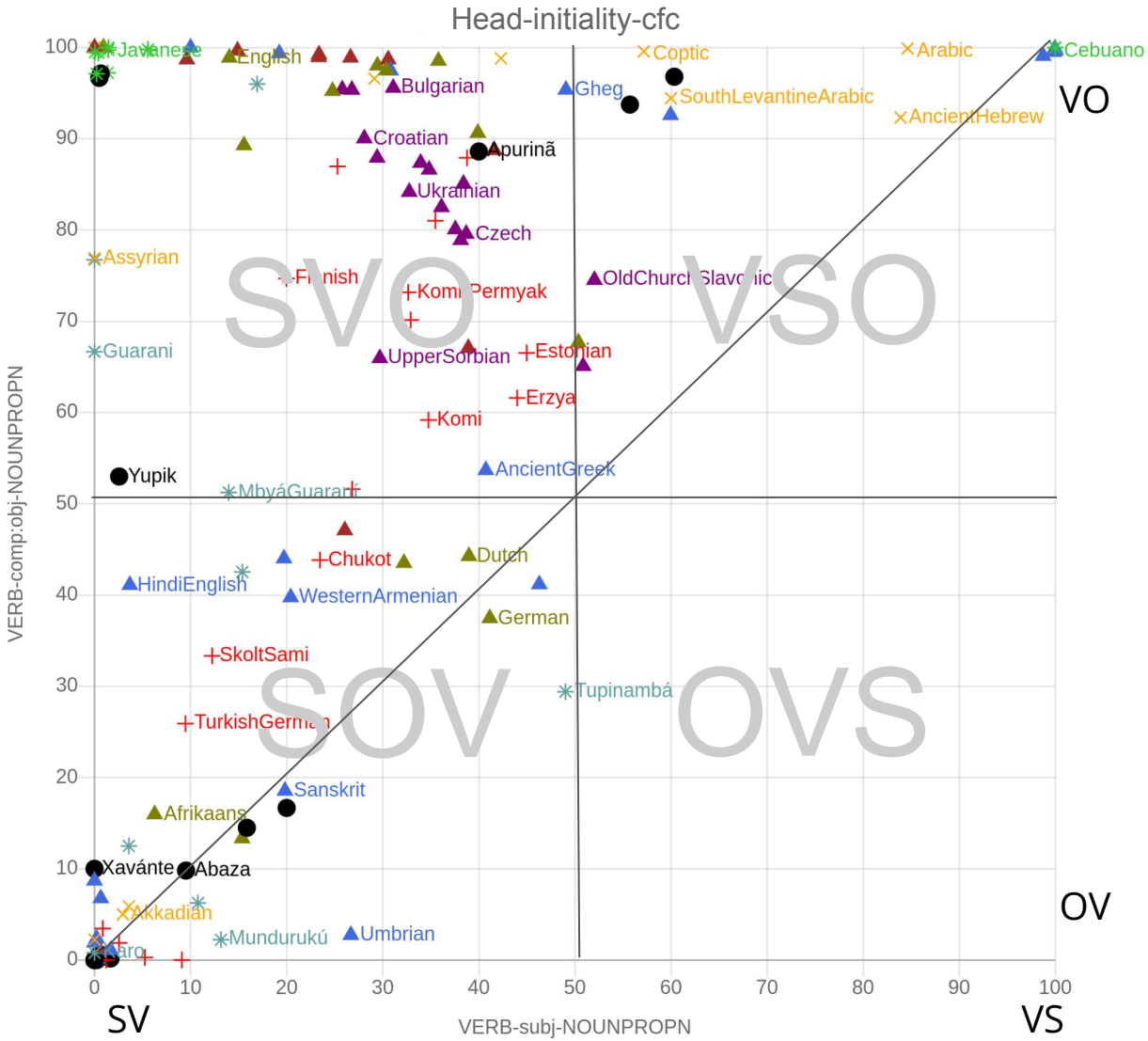
https://typometrics.elizia.net/

Tool: Typometrics

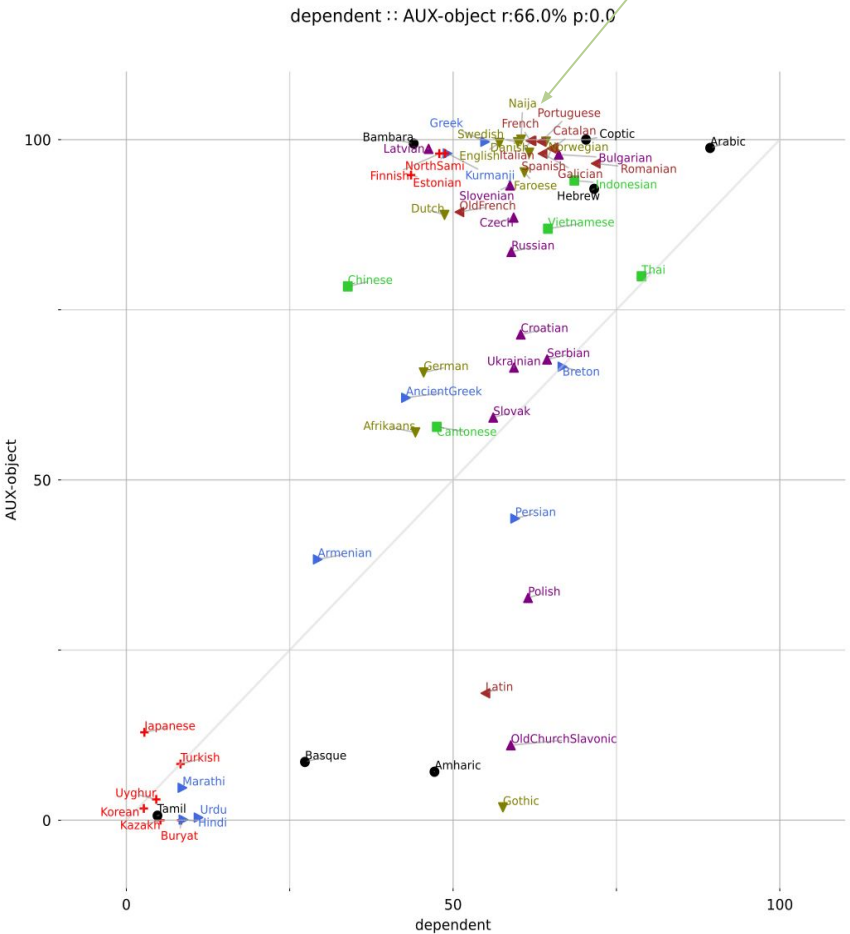
Quantitative statistical observation :

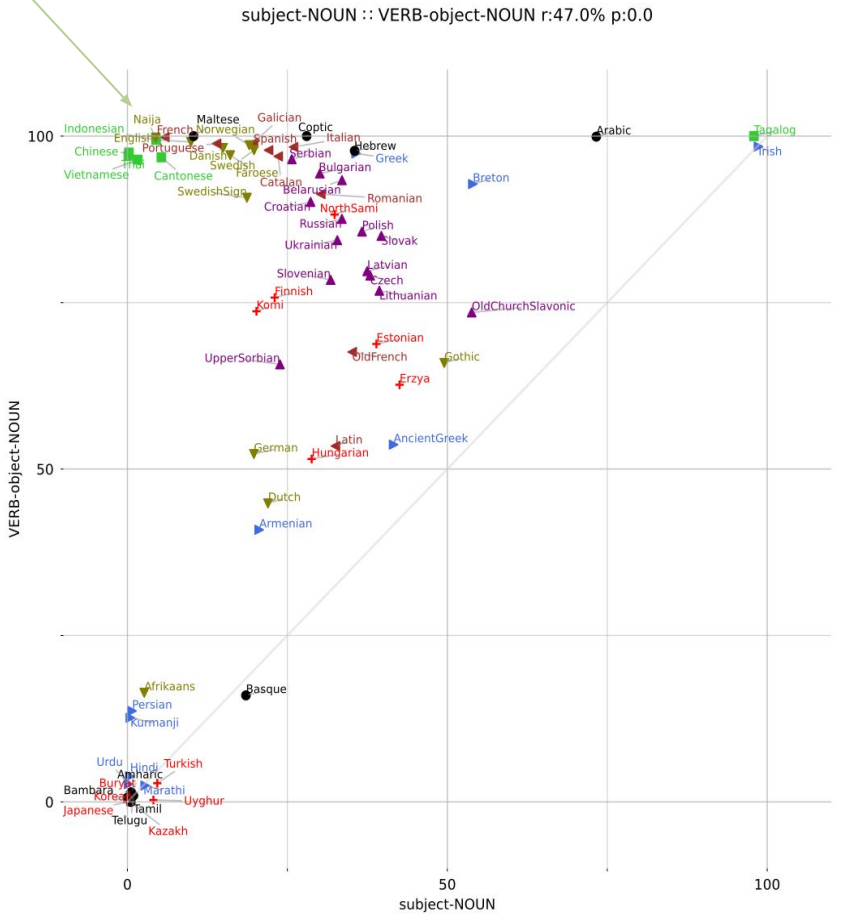
- For almost all the languages in our sample we have more objects on the right than subjects on the right.

https://typometrics.elizia.net/



Typometrics:
Naija, just another Germanic language?





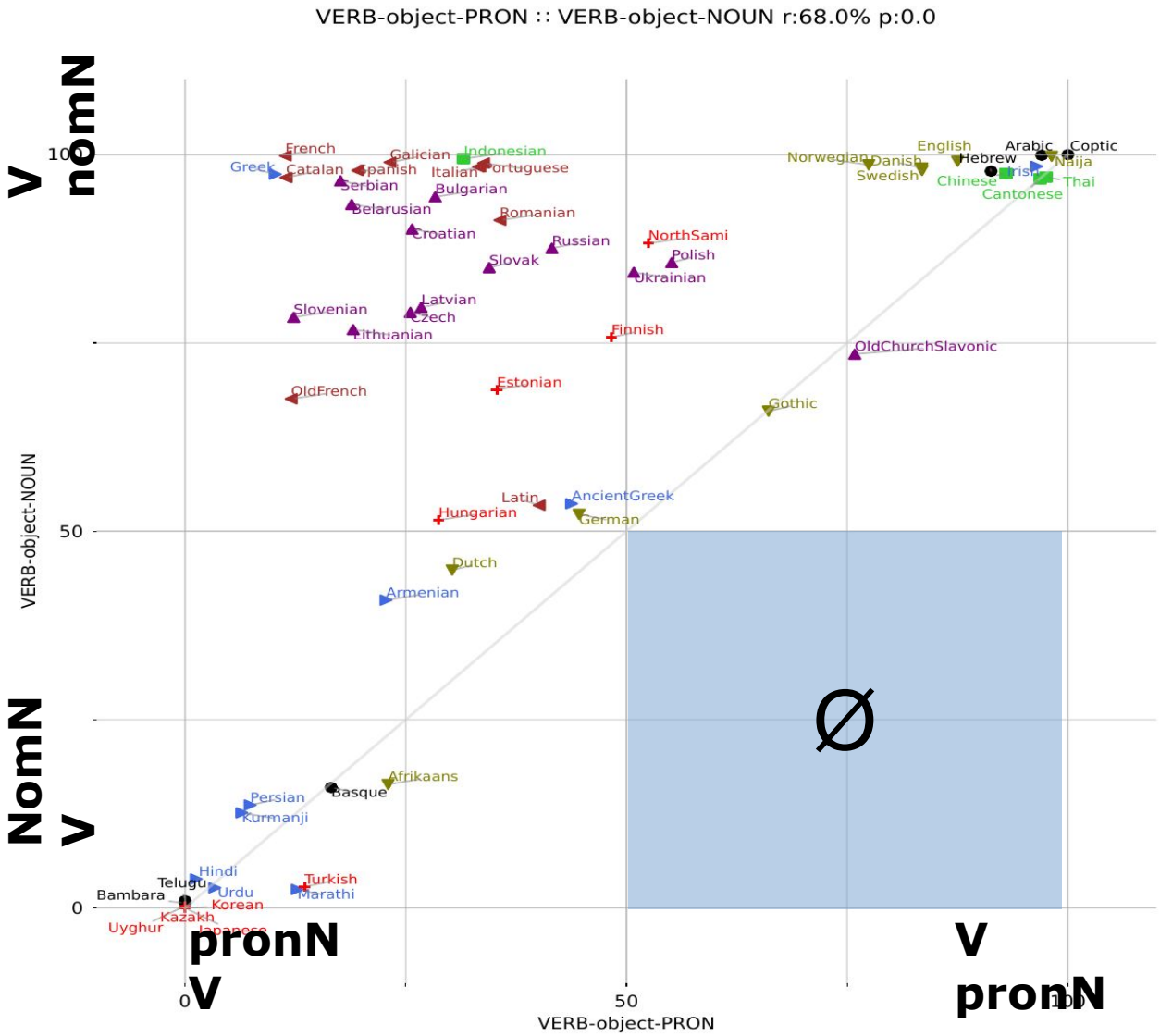
Typometrics: Tesnière's "typological classification of languages according to the nature of linearization"

- **Languages**
  - **centrifugal**
    - strict
      - Bantu languages
      - Semitic languages
      - (Hebrew, Arabic)
      - Austronesian languages
      - (Malaysian, Polynesian)
    - mitigated
      - Black-African languages
      - Tasmanian
      - Old Asian languages
      - (Elamite, Sumerian)
      - American languages
      - Basque
      - Celtic languages
      - Romance languages (Italian, Spanish, **French**)
      - Indo-European languages
        - Greek
        - Latin
        - Germanic languages
        - (English, **German**)
        - Slavic languages (Russian)
  - **centripetal**
    - mitigated
      - Languages of Australia
      - Papuan languages
      - Languages of Andaman islands
      - Sino-Tibetan languages
      - (Chinese)
    - strict
      - Caucasian languages
      - South-African languages
      - (Khoekhoe, Bushman)
      - Dravidian languages
      - Burushaski
      - Hyperborean languages
      - Ural-Altaic languages (including Japanese and Korean)

Greenberg's Universal 25

"If the pronominal object follows the verb, so does the nominal object"

V pronN → V N



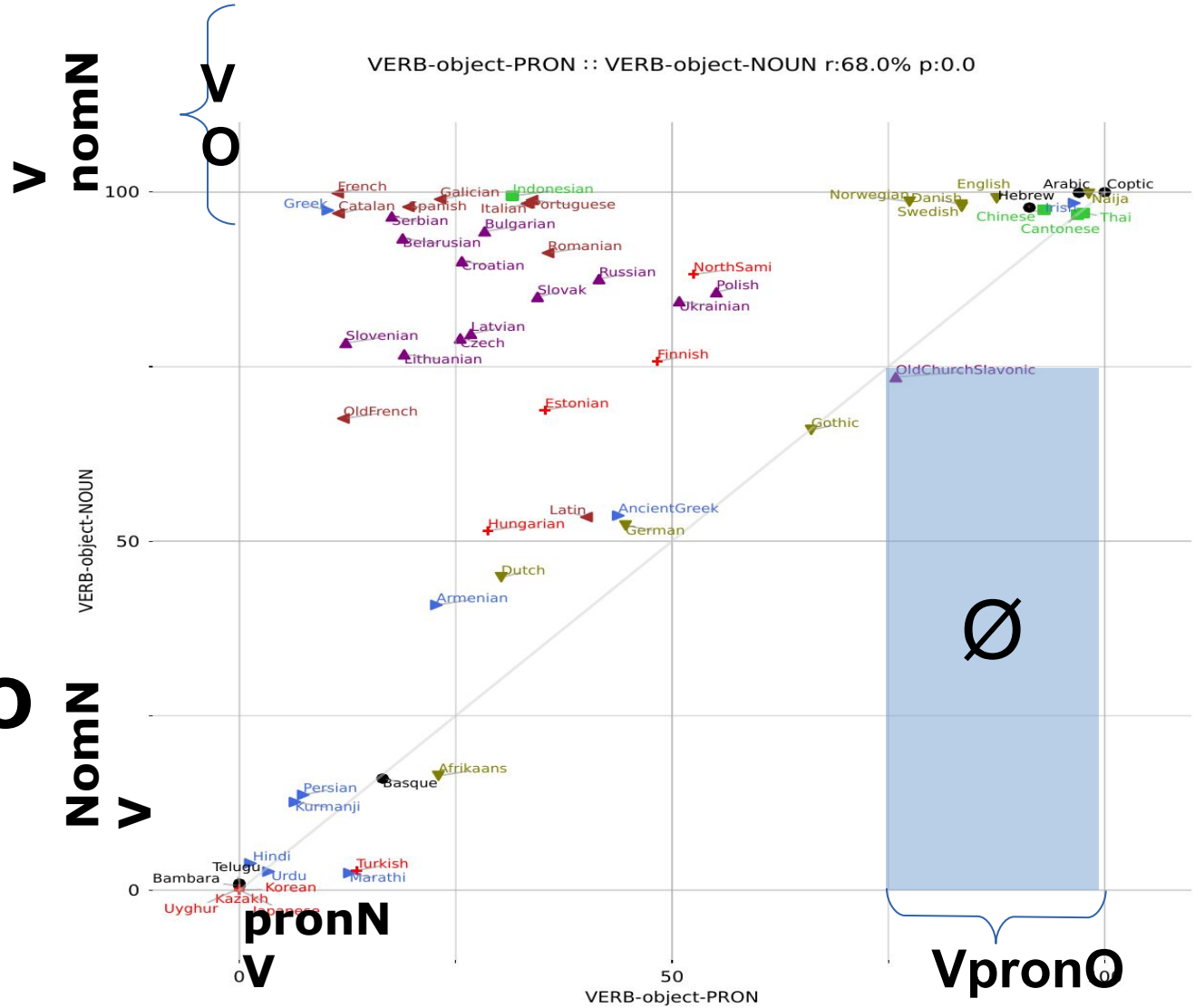
Universal 25

"If the pronominal object follows the verb, so does the nominal object"

for a certain *a*,

VpronO ≥ *a* → VO ≥ *a*

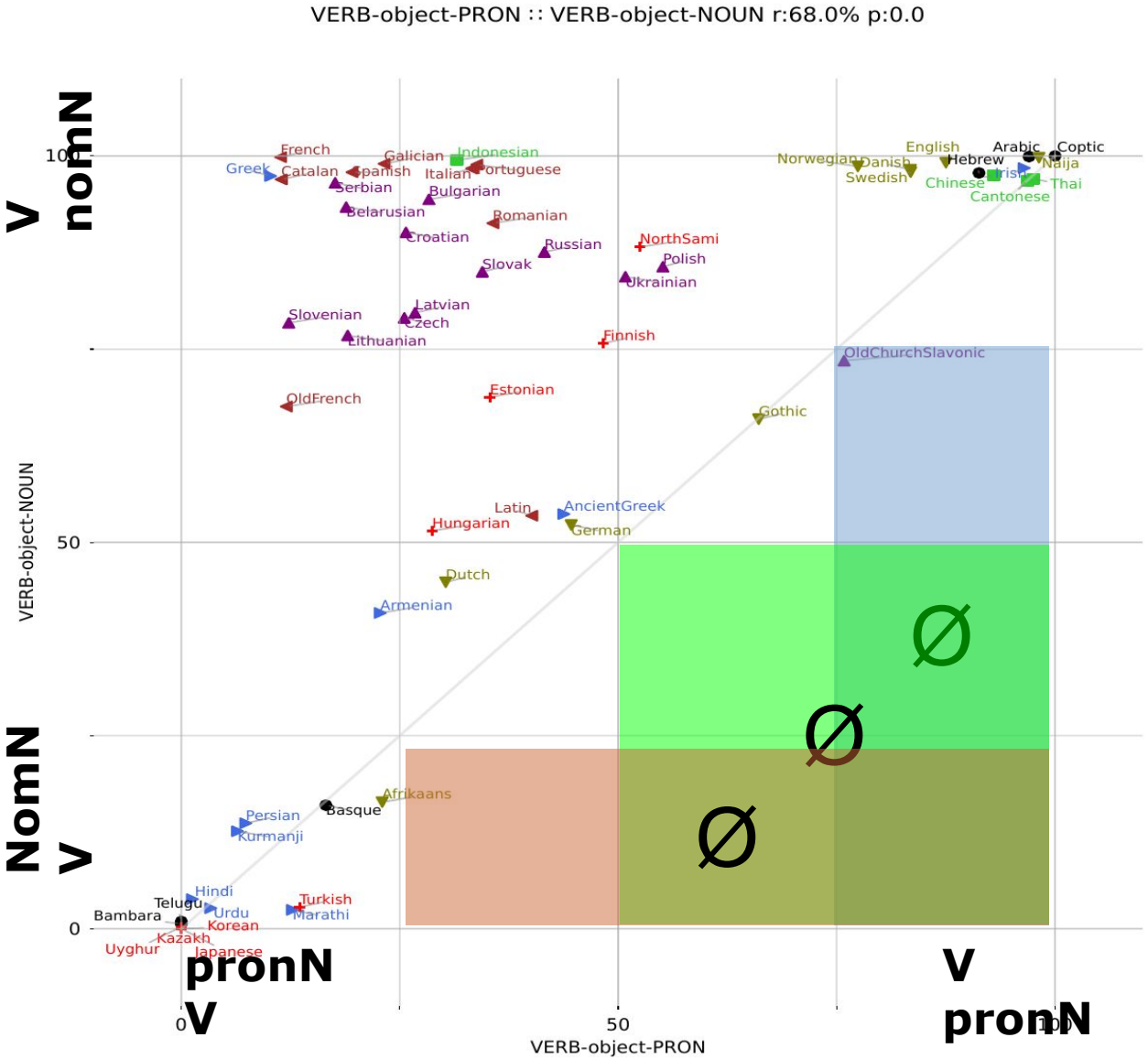
Which value to



Universal 25’

For each *a*,

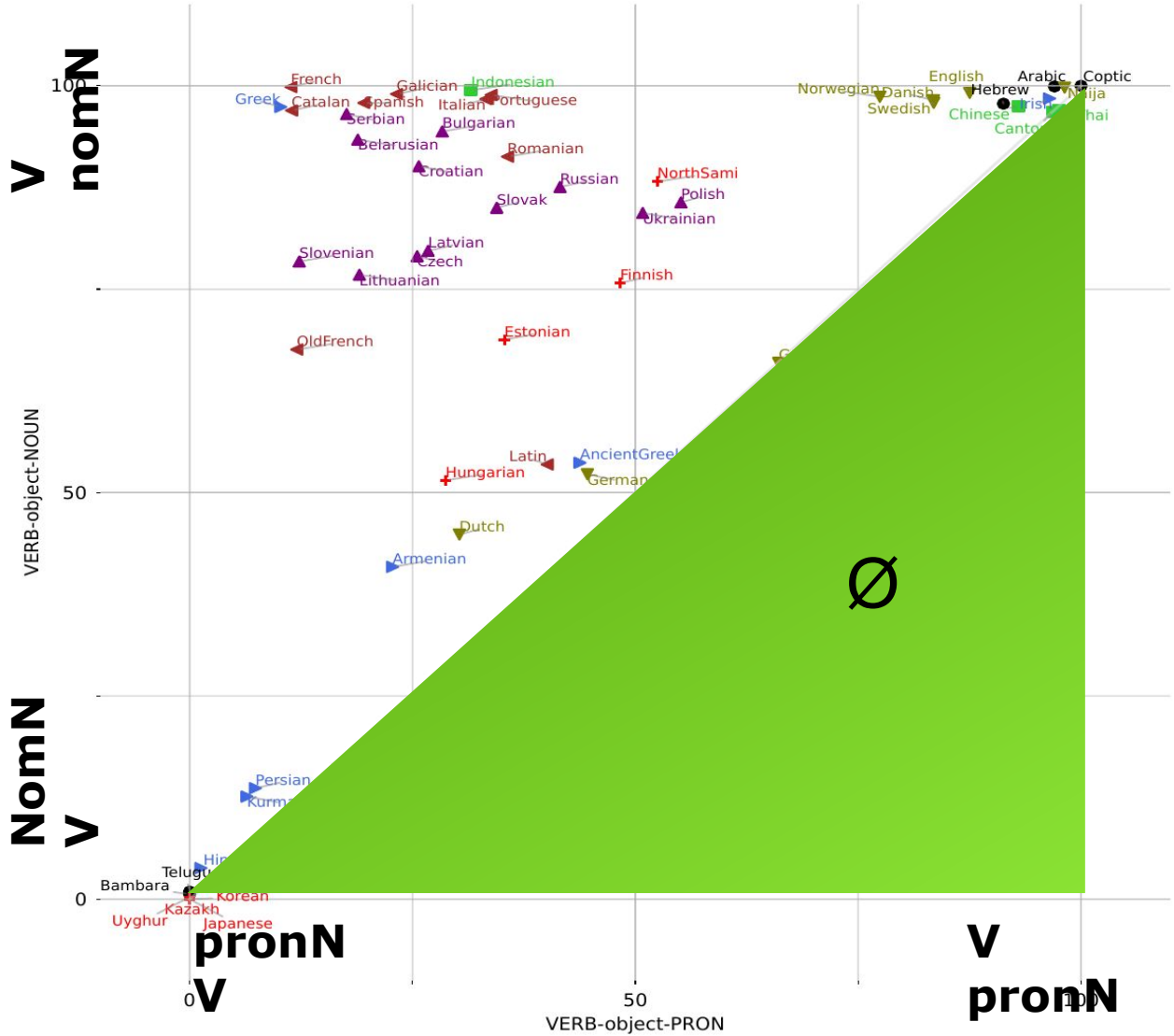
VpronO ≥ *a* → VO ≥ *a*



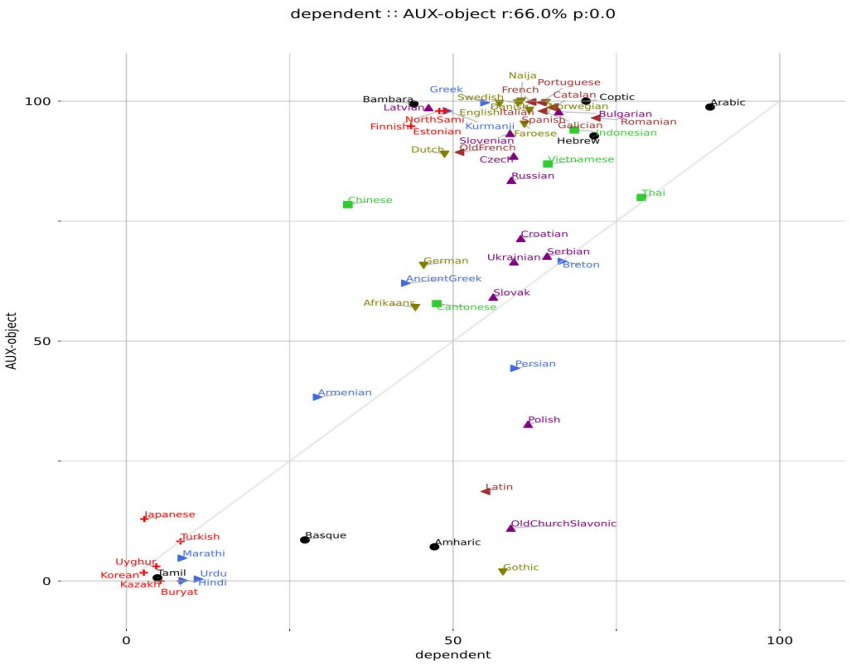
Universal 25’

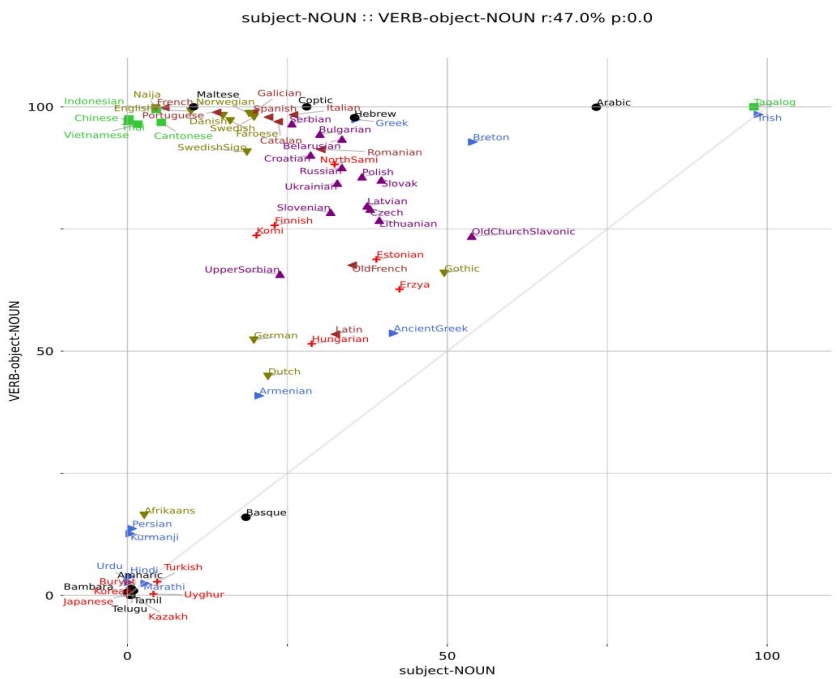
For each *a*,

VpronO ≥ *a* → VO ≥ *a*



https://typometrics.elizia.net/#/





4) Annotation

Annotation is interesting

- Because annotated corpora allow for
  - direct access to certain phenomena
  - different types of measures
- Because annotated corpora can be used for
  - Training automatic parsers
  - Economically interesting
- Because the annotation procedure in itself is linguistically interesting:
  - How to put linguistically coherent analyses on all types of real-world data?

ArboratorGrew

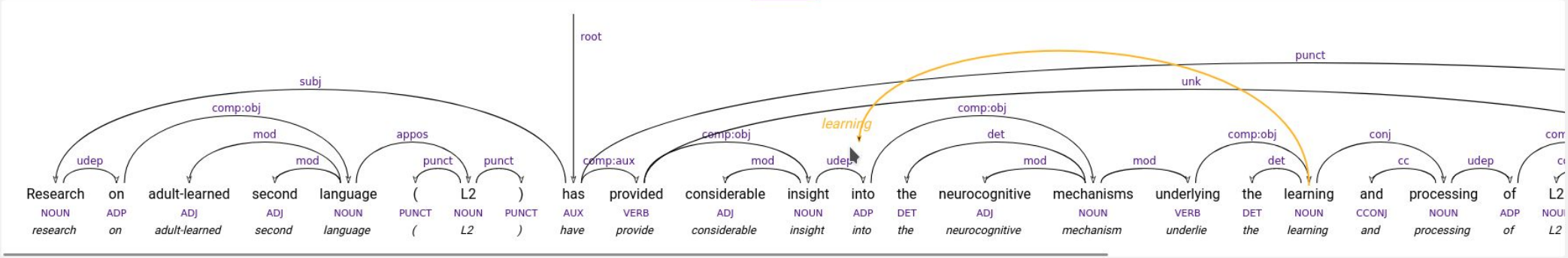
ARBORATOR grew / GUM SUD / en_gum-sud-dev

1 GUM_academic_exposure-1 Introduction

kmgrds

2 GUM_academic_exposure-2 Research on adult-learned second language (L2) has provided considerable insight into the neurocognitive mechanisms underlying the learning and processing of L2 grammar [1] – [11]

kmgrds



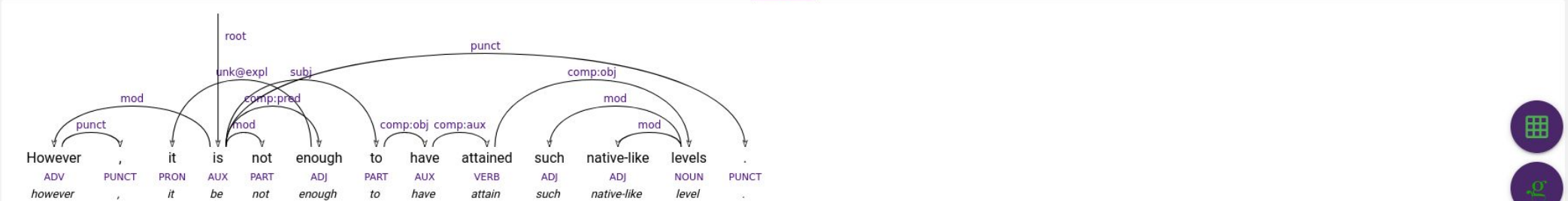
text_en

3 GUM_academic_exposure-3 Of interest here, studies suggest that, despite the difficulties in acquiring L2 grammar, adult learners can approximate native-like levels of use and neurocognitive processing [12] – [15].

kmgrds

4 GUM_academic_exposure-4 However, it is not enough to have attained such native-like levels.

kmgrds

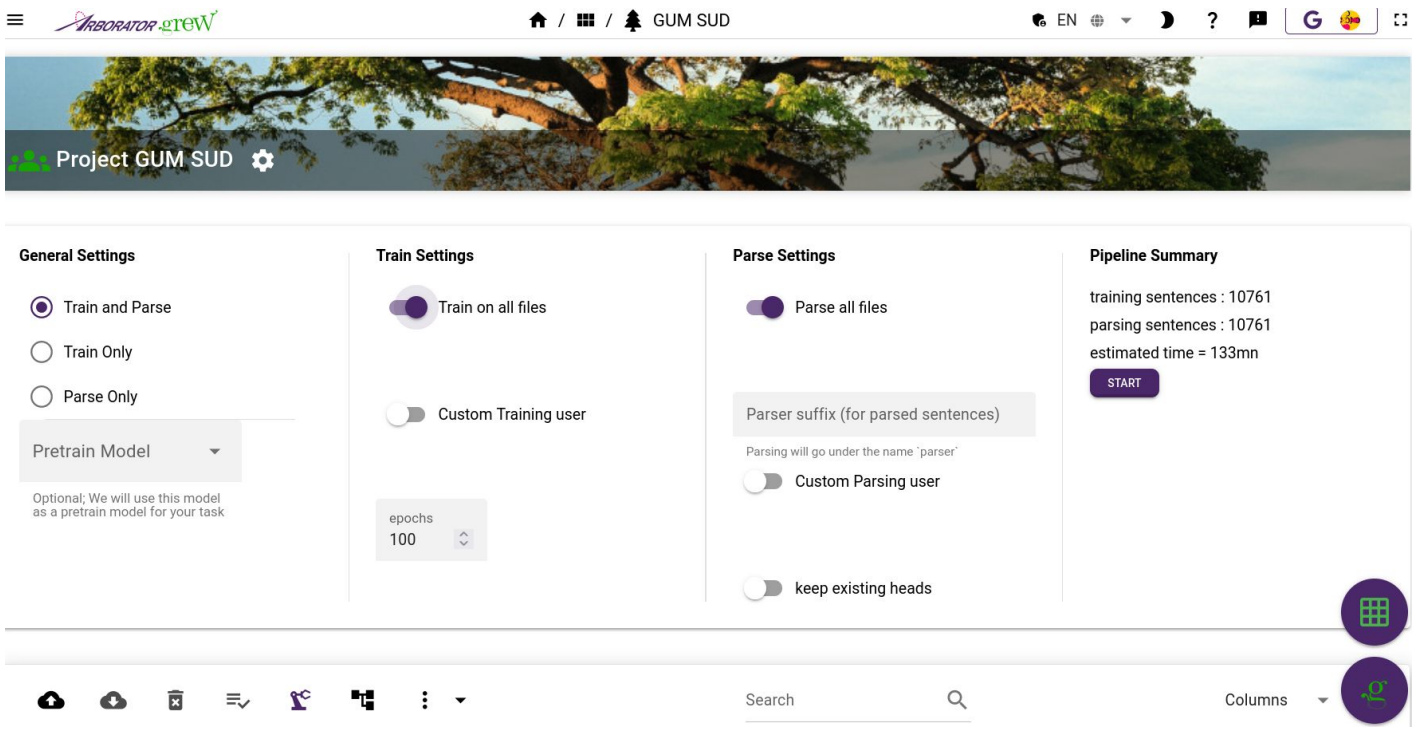


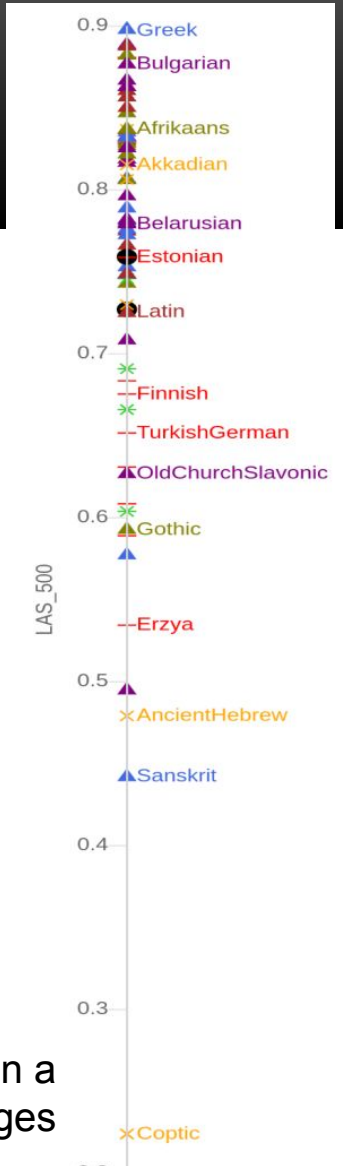
Some cool features of ArboratorGrew

- Collaborative annotation
- **Classroom annotation:**
  - **Exercise mode**
  - Search, description, quantification, correction of syntactic constructions
- GitHub synchronization
- Lexical view of the treebank
- Grew integration
- Parser bootstrapping
  - Annotate a few sentences of your new language, your new corpus, train a parser and pre-parse the rest of the corpus

→ *Gaudi ex cathedra*

Bootstrapping





Average performance of Trankit and UDify measured by LAS on a dataset of 500 sentences for each of the 69 languages

Let’s dive right in

https://arboratorgrew.elizia.net/#/projects/ESSLLI_2023_Treebanks/treebanking.GUM.test

Graph rewriting

grew

Hands-on session

First exercise: *How are adjectives analyzed in SUD?*

SUD_English-GUM@2.12

1. All adjectives
   - 1000 occurrences in 8.42%
   - Find regularities
2. Adjectives linked to a NOUN with the `mod` relation
   - 1000 occurrences in 46.64%
3. Adjectives linked to an AUX with the `comp:pred` relation
   - 1000 occurrences in 83.61%
4. Remaining cases

1

```
pattern {
  A[upos=ADJ]
}
```

2

```
pattern {
  A[upos=ADJ];
  N[upos=NOUN];
  N -[mod]-> A;
}
```

3

```
pattern {
  A[upos=ADJ];
  V[upos=AUX];
  V -[comp:pred]-> A;
}
```

4

```
pattern { A[upos=ADJ] }
without { N[upos=NOUN]; N -[mod]-> A }
without { V[upos=AUX]; V -[comp:pred]-> A }
```

Grew

- http://universal.grew.fr/
- SUD_English-GUM

```
pattern { N [form="ants"] }
pattern { N [lemma="ant"] }
pattern { N [upos="NUM"] }
pattern { GOV -[comp:obj]-> DEP }
pattern {
  GOV [upos = VERB];
  GOV -[subj]-> DEP;
}
```

Grew

- Clustering
  - GOV.upos
  - GOV.form
- Find the
  - most common verb, auxiliary, discourse
  - most common subject category
  - most common word after “what the”
  - most common word

POS?

- Use UD's Universal POS tags
  - Possibly with extra features
  - except if you have good reasons to do otherwise
- Open class words: ADJ ADV INTJ NOUN PROPN VERB
- Closed class words: ADP AUX CCONJ DET NUM PART PRON SCONJ
- Other: PUNCT SYM X

Micro-syntactic relation

- **subj**, for subjects
- **mod** for modifiers
- **comp:obj**, for direct objects
- **comp:obl**, for oblique complements
- **comp:pred,** for predicative complements
- **comp:aux,** for relations between a TAM (Time/Aspect/Modality) auxiliary and the full verb
- **comp:cleft**, for the cleft-clauses
- **compound:svc**, for serial verb constructions

List relations

- **conj** for paradigmatic lists
  - **conj:coord**, for coordination (conjuncts have different referents): *Mary and Peter*
  - **conj:dicto**, for disfluency and reformulation (different denotations of a same referent):
    the g- the girl the little girl
  - **conj:appos**, for apposition (two predications on a same referent): *Mary, my best friend, …*

Macro-syntactic relations

- **discourse**, for discourse markers:
  well, er, you know, isn't it?
- **dislocated**, for dislocated phrase:
  Peter, I know him very well
- **vocative**, for adresses:
  Peter, what are you doing?
- **parataxis:parenth**, for parenthetical
- **punct**, for punctuation

Grew

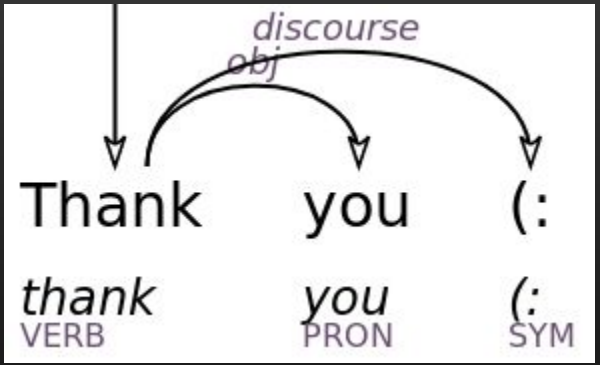
- What's the most common proper noun?
- What's the most common subject?
- What's the most common subject that is not a pronoun?
- How many appositions exist between two different categories?
- How many VERBs have a determining dependent?
- What's the most common verbal lemma that has a subject but no object?
- If a verb has "years" as object, what's its most common subject?

Grew

Speech and Language Processing (OPT5)
2024_SLP2024

https://ecampus.paris-saclay.fr/course/view.php?id=154814

Quiz: https://ecampus.paris-saclay.fr/mod/quiz/view.php?id=1741667


discourse
obj
Thank
you
(:
thank
you
(:
VERB
PRON
SYM