Dependency syntax
SUD vs UD annotation schemes

Sylvain Kahane
(Modyco, Paris Nanterre & CNRS / IUF)

Chișinău, July 9, 2024

Objectives

- Principles of dependency syntax
- Annotation scheme: UD and SUD
  - general principles
  - UD and SUD tag sets
  - conversion between UD and SUD
- Practical annotation (with ArboratorGrew)
  - start a project, manual annotation, annotation with Grew rules
  - annotation in groups for various languages

Plan

- Class 0 (Monday, session 6)
  - initiation to Grew-match
  - start a project on ArboratorGrew
- Class 1 (Tuesday, session 7)
  - Dependency trees
  - History of syntactic representations and treebanks
  - What to do with treebanks?

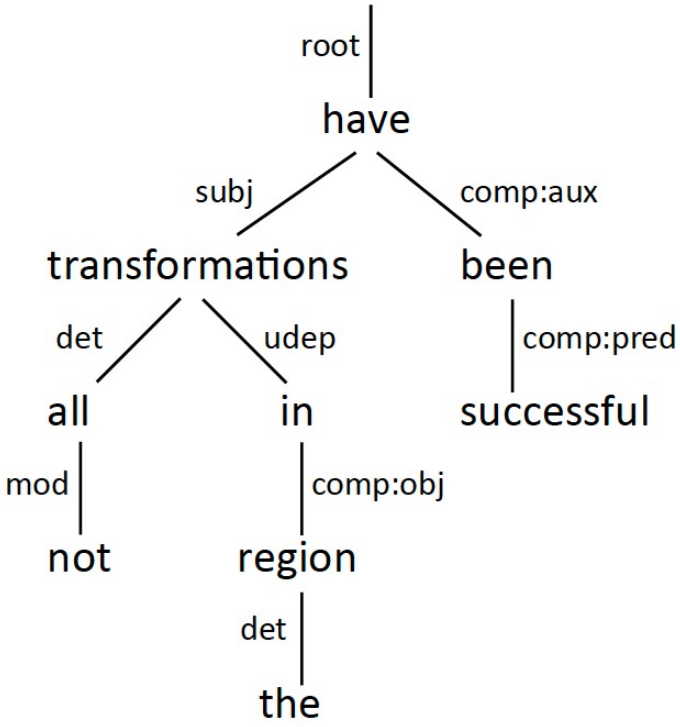
Plan

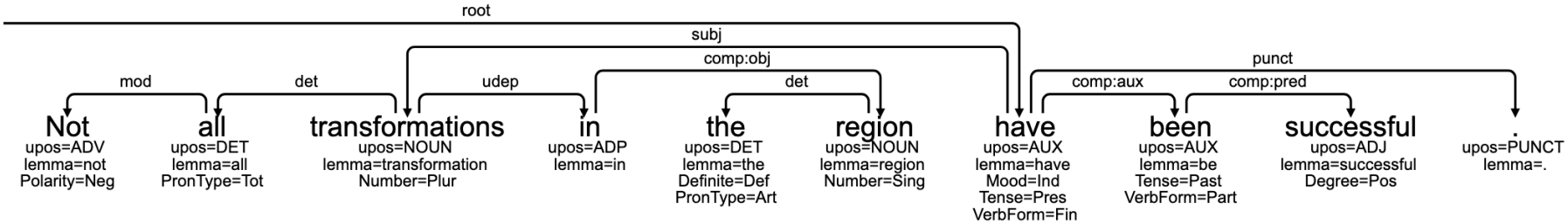
- Class 2 (Wednesday, session 12)
  - Definition of the syntactic structure
    - Exercises
  - SUD and UD annotation scheme, conversion
- Class 3 (Thursday morning, session 15)
  - Annotation of the participants' data
  - More on ArboratorGrew (Bruno)
- Class 4 (Thursday afternoon, session 17)
  - More on SUD and UD annotation scheme
  - Annotation of the participants' data

Dependency trees

Dependency tree

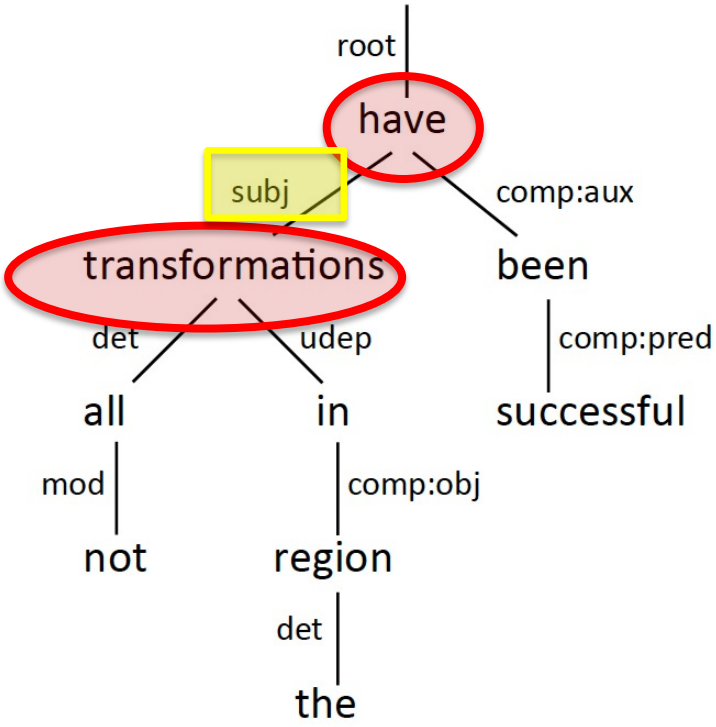
- The syntactic structure can be represented by a dependency tree (Beauzée 1767, Tesnière 1934, 1959, Hudson 1984, 2010, Mel'čuk 1988)





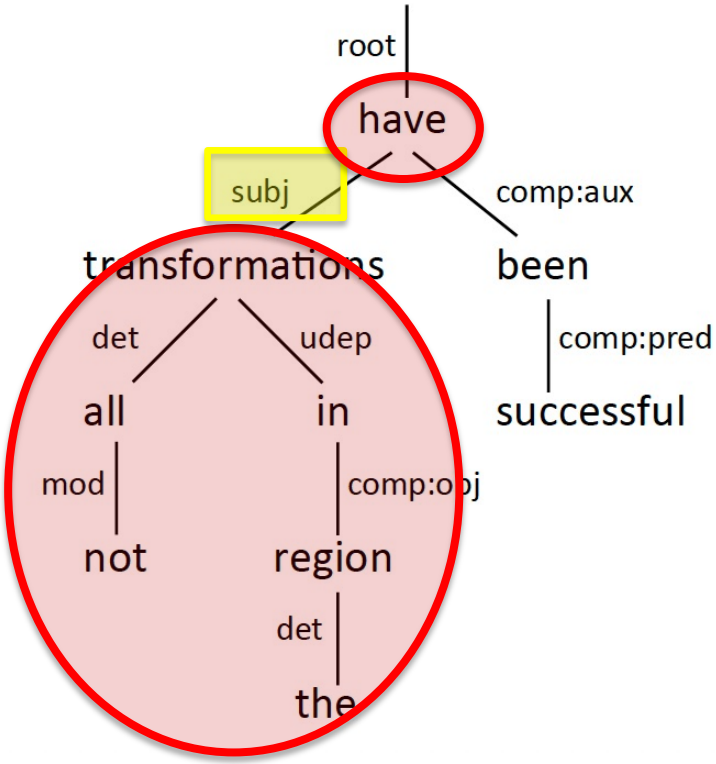
How to read a dependency tree

- Words are linked by labeled dependencies



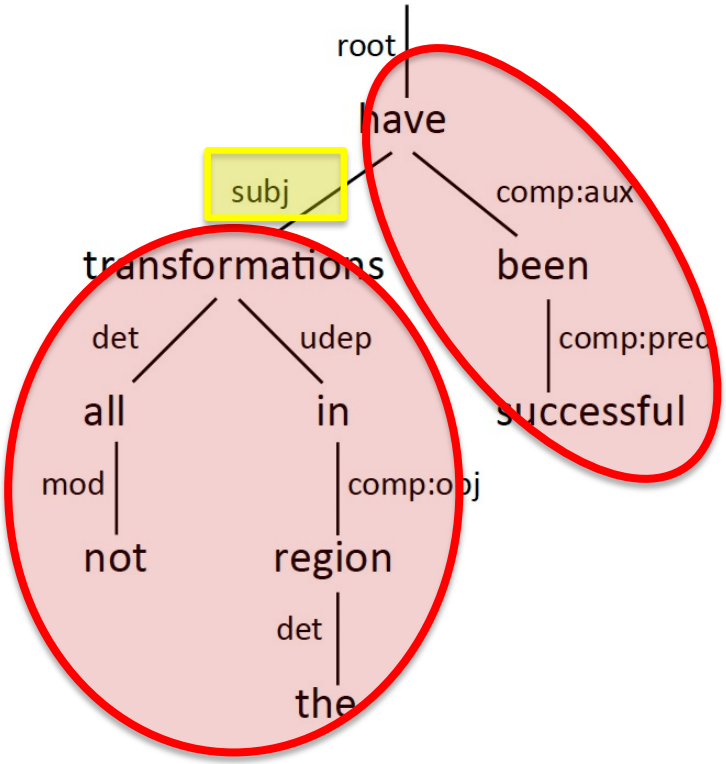
How to read a dependency tree

- combination with the projection
- projection of W = phrase formed by all the words dominated by W



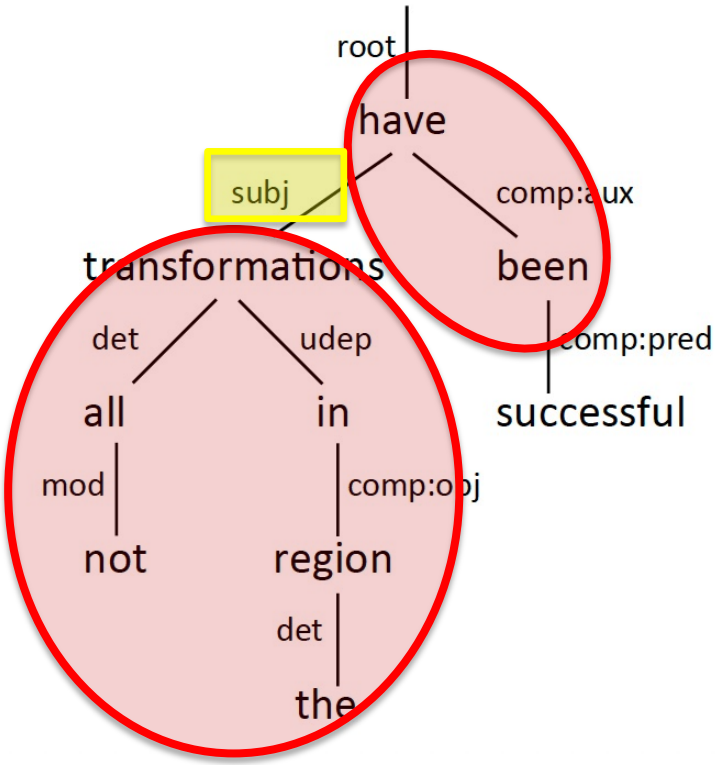
How to read a dependency tree

- Another combination:
  S -> NP VP



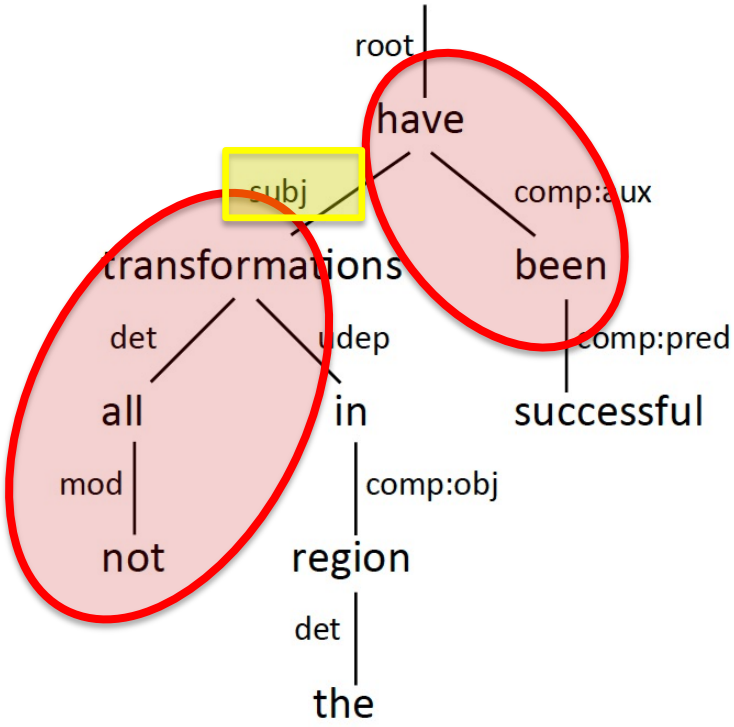
How to read a dependency tree

- every combination linking a unit on one side with a unit on the other side (not necessary words)



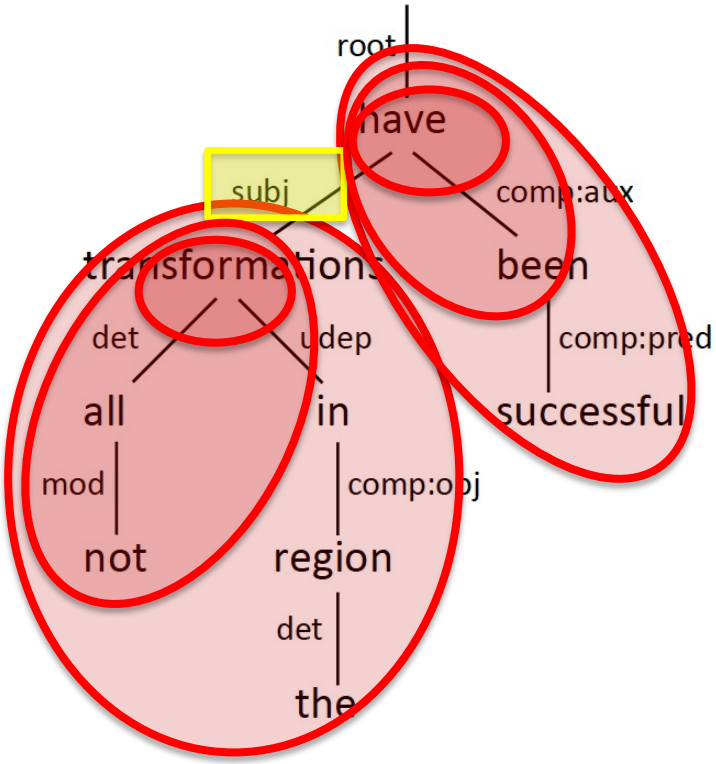
How to read a dependency tree

- every combination linking a unit on one side with a unit on the other side (not necessary words)



How to read a dependency treeHow to read a dependencv tree

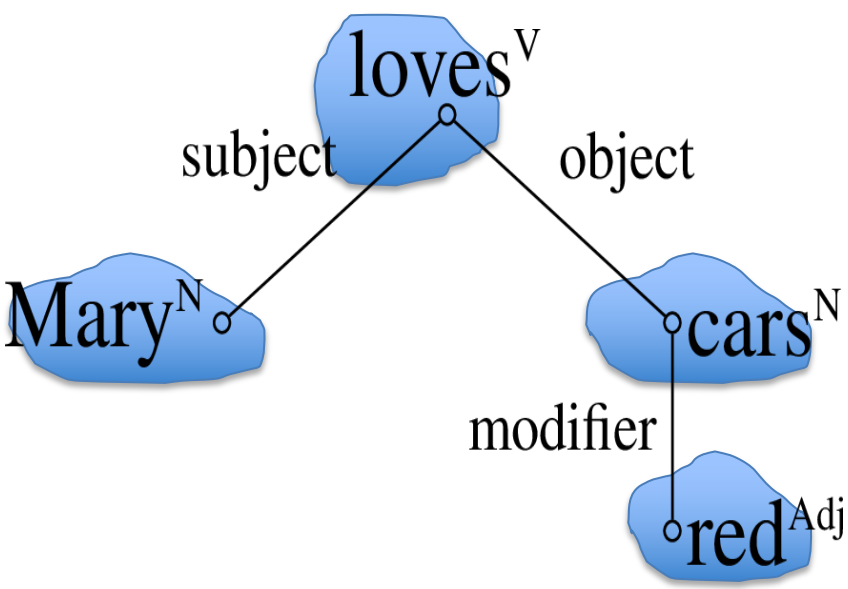
- a dependency represents a set of combinations
- these combinations are said equivalent

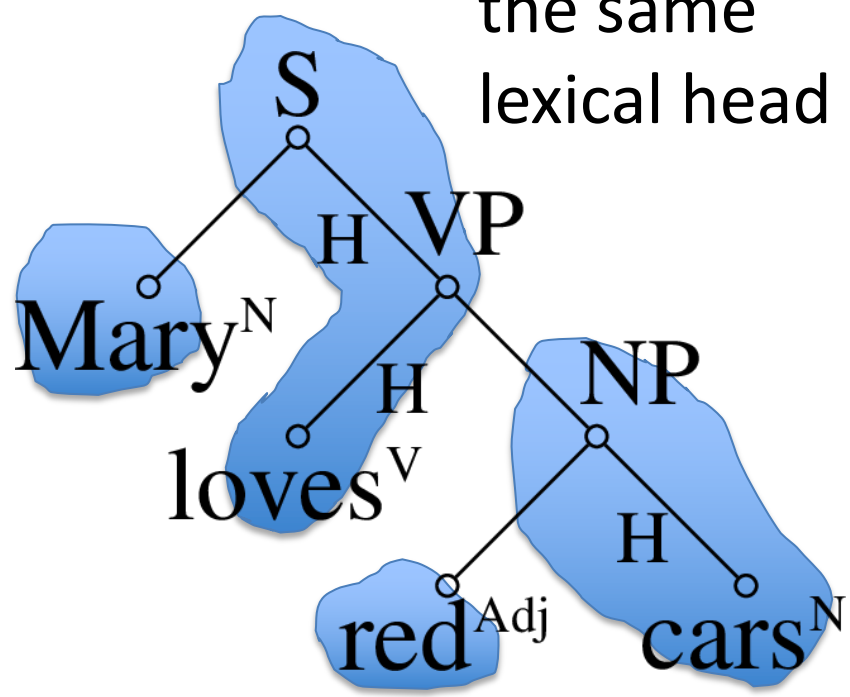


Equivalence with phrase structure trees

- an headed constituency tree subsumes a depencency tree (Lecerf 1961)

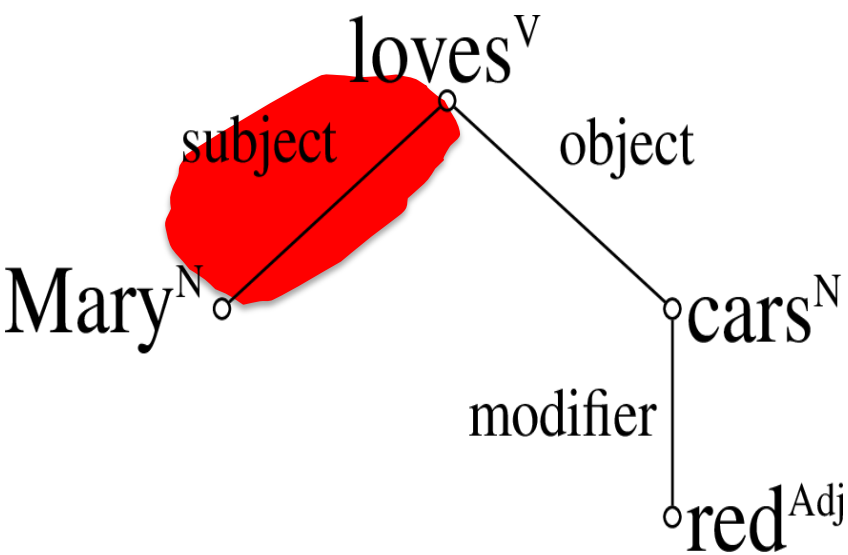
by collapsing nodes with the same lexical head

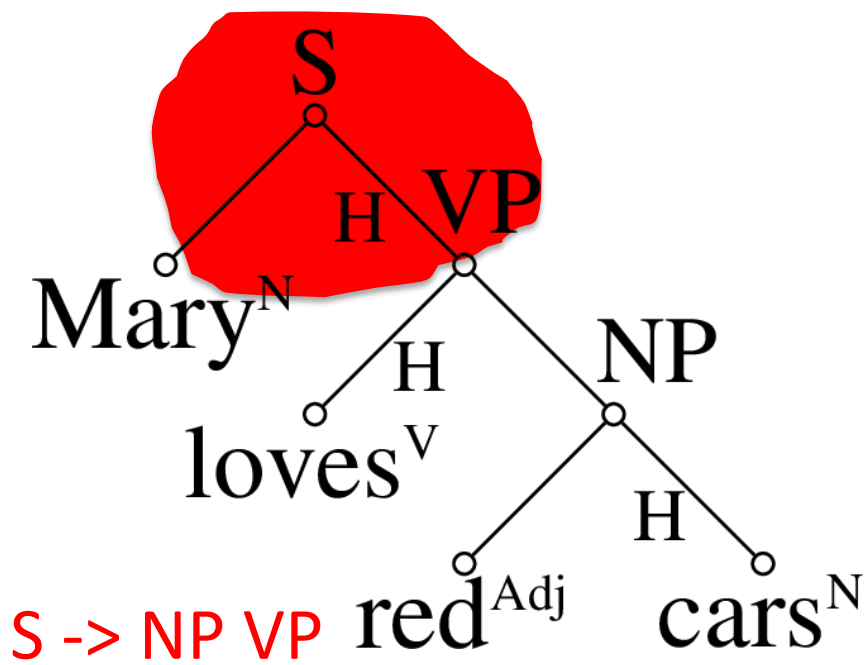




Equivalence with phrase structure trees

- a depencency tree subsummes a headed constituency tree (without linear order)

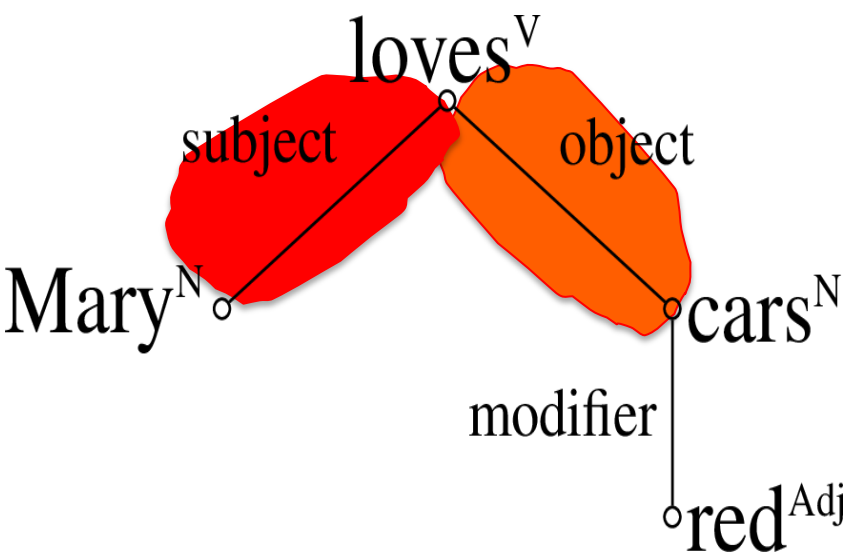


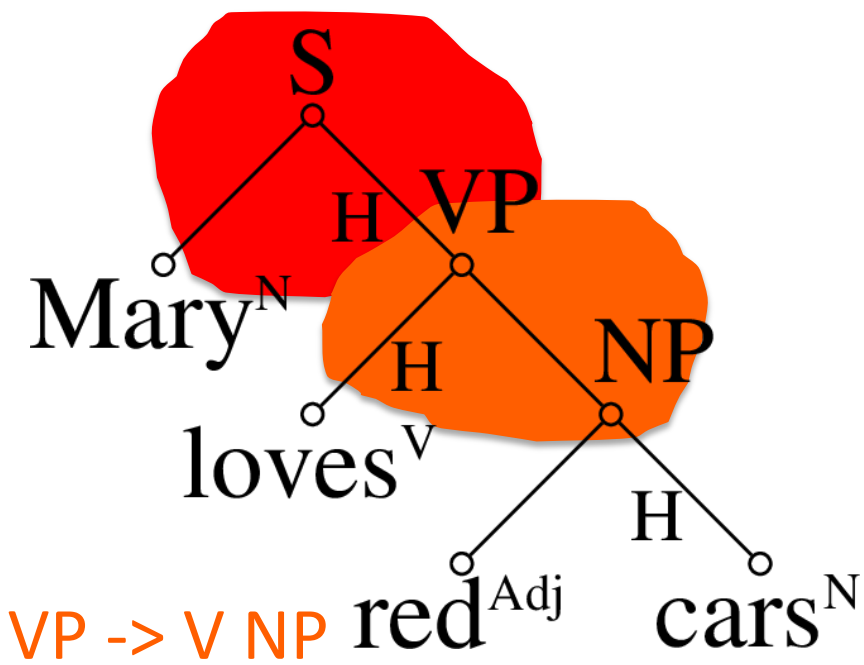


S -> NP VP

Equivalence with phrase structure trees

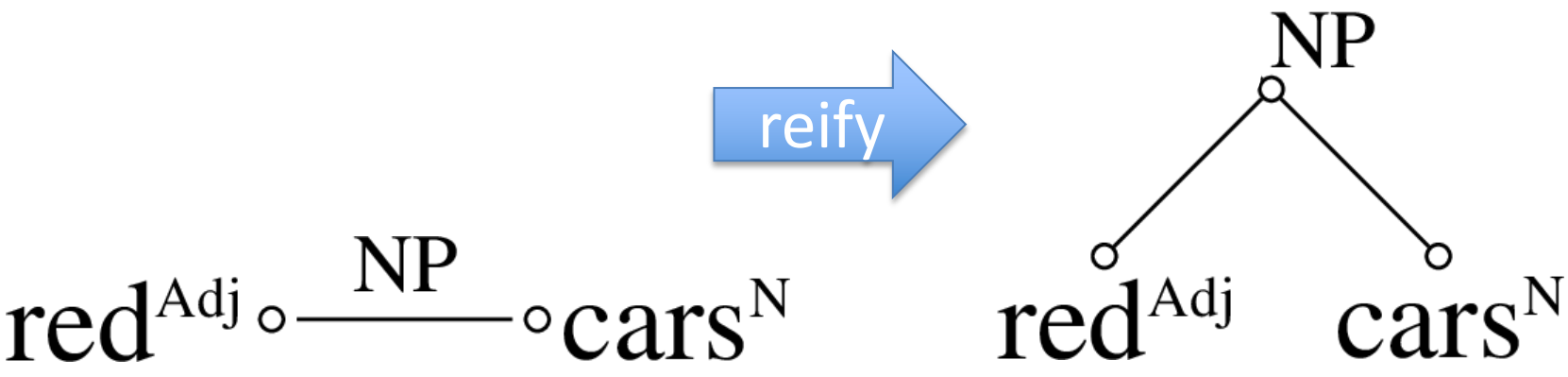
- a depencency tree subsummes a headed constituency tree (without linear order)





Reification

- the relation between an edge and its vertices become an edge



dependency
= binary link

binary-branching tree

Properties of a tree

- a directed graph is a tree if and only if
  - all vertices are connected
  - every vertex has a unique governor, except the root

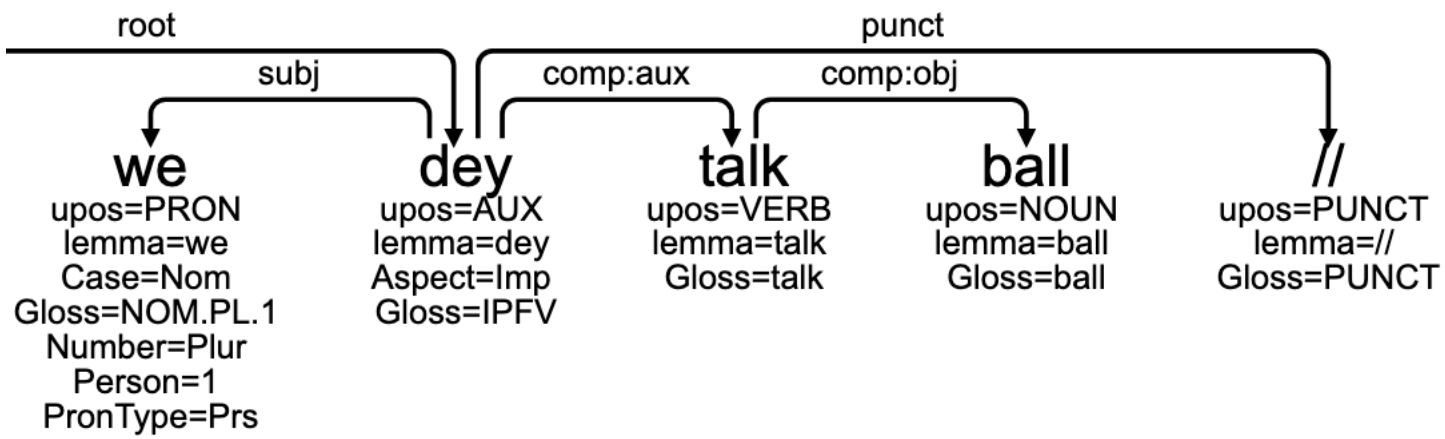
Properties of a tree

- a directed graph is a tree if and only if
  - all vertices are connected
  - every vertex has a **unique governor**, except the root

=> conll format

Conllu (tabular format)

# speaker_sex = M							
# text = we dey talk ball //							
# text_en = We discuss football.							
1	we	we	PRON	_	Case=Nom|Number=Plur|Person=1|PronType=Prs	2	subj
2	dey	dey	AUX	_	Aspect=Imp	0	root
3	talk	talk	VERB	_	_	2	comp:aux
4	ball	ball	NOUN	_	_	3	comp:obj
5	//	//	PUNCT	_	_	2	punct



Syntactic diagrams’ history

Claude Bufffier (1709)

- *Un homme qui étourdit les gens qu'il rencontre avec de frivoles discours, a coutume de causer beaucoup d'ennui à tout le monde*. Je dis que dans ce discours, tous les mots sont pour modifier le nom *un homme*, & le verbe *a coutume*, & que c'est en cela que consiste tout le mistére & toute l'essence de la sintaxe des langues :
  - 1° le nom *un homme*, est modifié d'abord par le *qui* déterminatif : car il ne s'agit pas ici d'un homme en général, mais d'*un homme* marqué & déterminé en particulier par l'action qu'il fait d'*étourdir* ;
  - de même il ne s'agit pas d'un homme *qui étourdit* en général, mais *qui étourdit* en particulier *les gens*, & non pas les gens en général, mais en particulier les gens *qu'il rencontre*.
  - Or cet homme qui étourdit ceux qu'il rencontre, est encore particularisé par *avec des discours*, & *discours* est encore particularisé par *frivoles*.
  - On peut voir le même dans la suite de la phrase : *a coutume* est particularisé par *de causer*, *de causer* est particularisé par ses deux régimes, par son régime absolu, savoir, *beaucoup d'ennui*, & par son régime respectif, *à tout le monde*.

  Voilà donc comment tous les mots d'une phrase quelque longue qu'elle soit, ne sont que pour modifier le nom & le verbe.

Claude Bufffier (1709)

- *A man who stuns the people he meets with frivolous speeches, is wont to cause a great deal of trouble to everyone*. I say that in this speech, all the words are to modify the noun *a man*, & the verb *is wont*, & that it is in this that consists all the mistery & all the essence of the sintax of languages:
  - 1° the noun *a man*, is first modified by the determinative *who*: for it is not a question here of a man in general, but of *a man* marked & determined in particular by the action he does of *stunning* ;
  - in the same way we are not talking about a man *who stuns* in general, but *who* particularly *stuns people*, & not people in general, but in particular *the people he meets*.
  - Now this man who stuns those he meets, is further particularized by *with speech*, & *speech* is further particularized by *frivolous*.
  - We can see the same in the rest of the sentence: *is wont* is particularized by *to cause*, *to cause* is particularized by its two regimes, by its absolute regime, namely, beaucoup d'ennui, & by its respective regime, *to everyone*.

So that's how all the words in a sentence, however long it may be, are only to modify the noun & the verb.

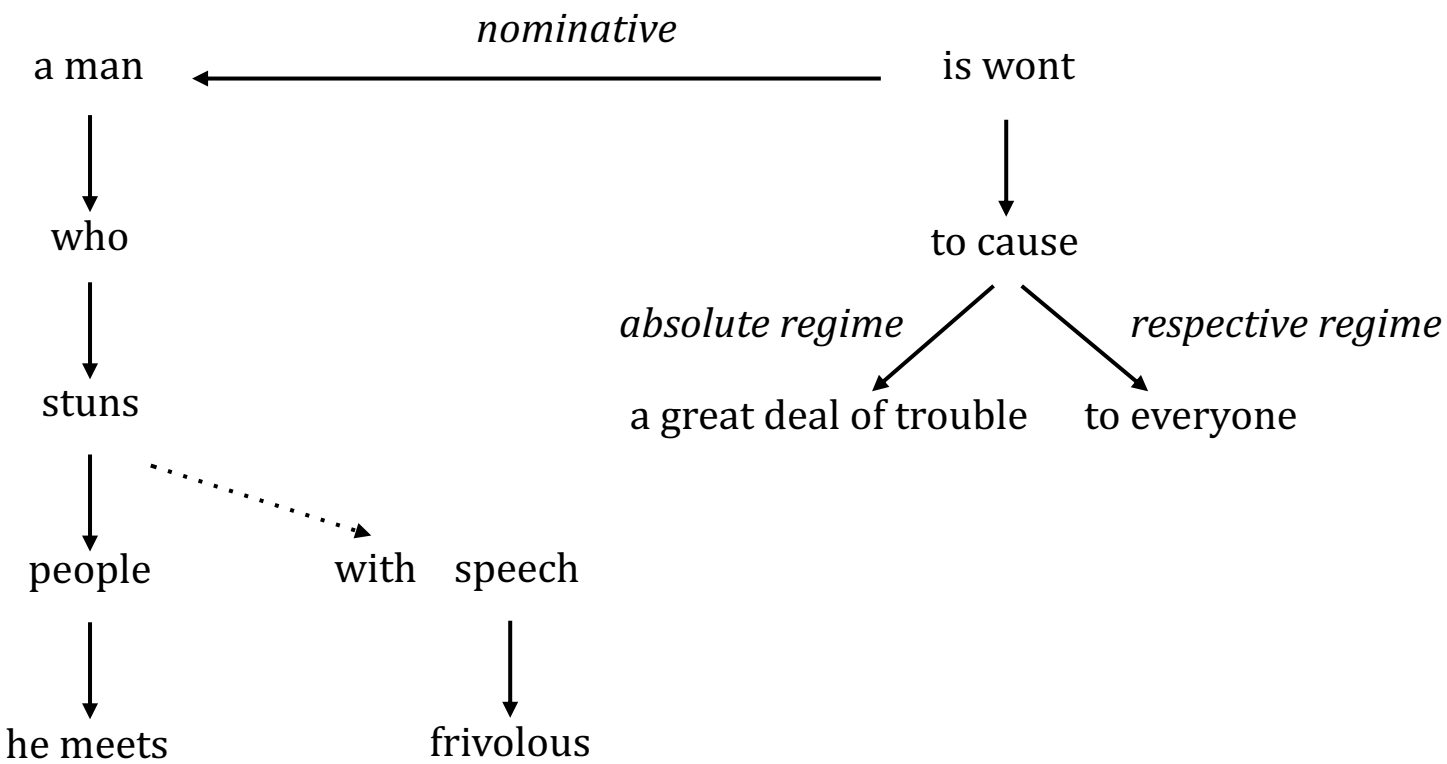
Claude Bufffier (1709)

- *A man who stuns the people he meets with frivolous speeches, is wont to cause a great deal of trouble to everyone*. I say that in this speech, all the words are to **modify** the noun *a man*, & the verb *is wont*, & that it is in this that consists all the mistery & all the essence of the sintax of languages:
  - 1° the noun *a man*, is first **modified** by the determinative *who*: for it is not a question here of a man in general, but of *a man* marked & **determined** in particular by the action he does of *stunning* ;
  - in the same way we are not talking about a man *who stuns* in general, but *who* **particularly** *stuns people*, & not people in general, but **in particular** *the people he meets*.
  - Now this man who stuns those he meets, is further **particularized** by *with speech*, & *speech* is further particularized by *frivolous*.
  - We can see the same in the rest of the sentence: *is wont* is **particularized** by *to cause*, *to cause* is **particularized** by its two regimes, by its **absolute regime**, namely, beaucoup d'ennui, & by its **respective regime**, *to everyone*.

So that's how all the words in a sentence, however long it may be, are only to **modify** the noun & the verb.

Claude Buffier (1709)

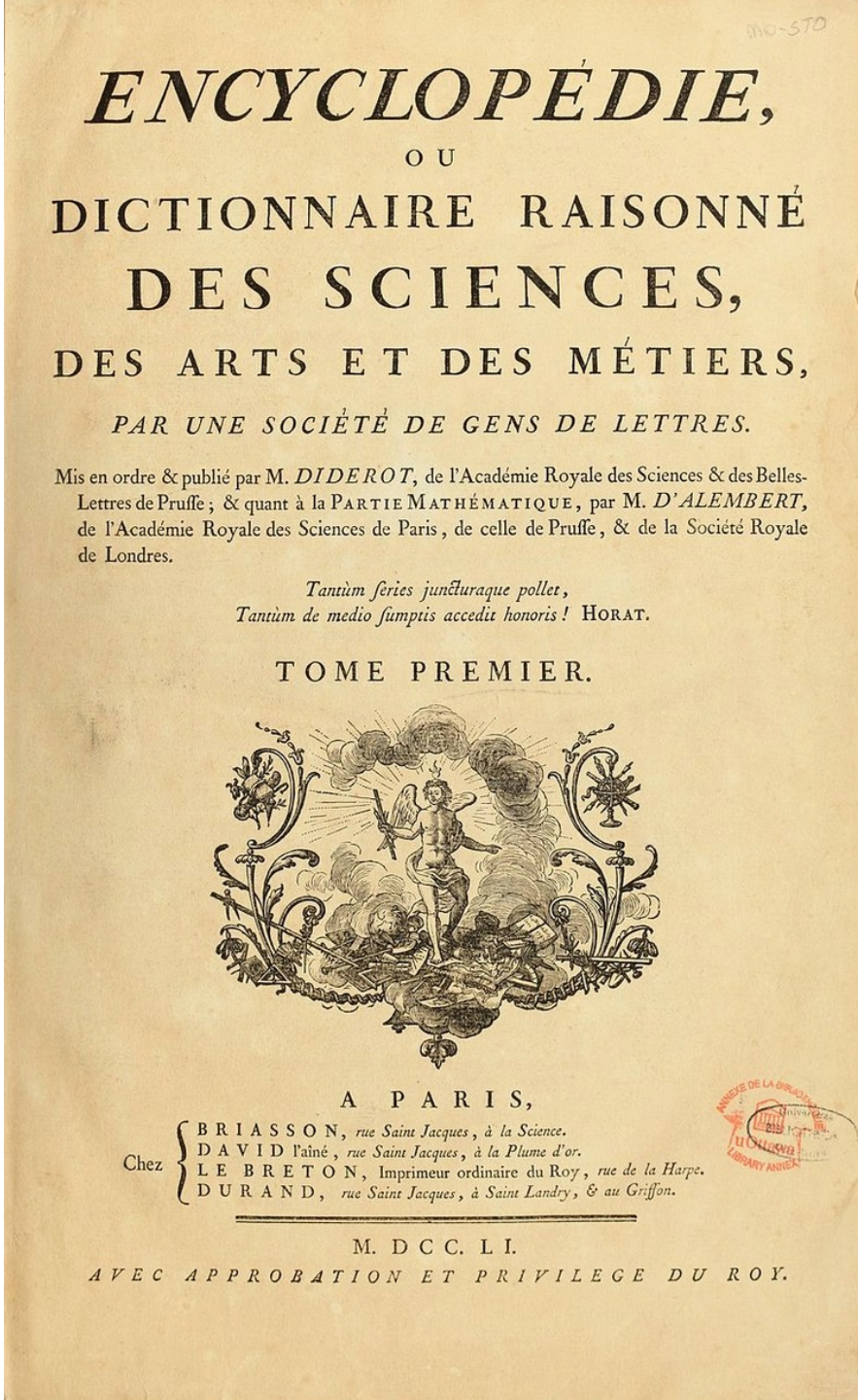
- proposed diagram according to Bufffier



Encyclopedia

or
Reasoned dictionary of sciences, arts and craft

- **Denis Diderot et Jean le Rond d'Alembert**
- 1751 — 1772



ENCYCLOPÉDIE,
OU
DICTIONNAIRE RAISONNÉ
DES SCIENCES,
DES ARTS ET DES MÉTIERS,
PAR UNE SOCIÉTÉ DE GENS DE LETTRES.

Mis en ordre & publié par M. DIDEROT, de l'Académie Royale des Sciences & des Belles-Lettres de Pruſſe; & quant à la PARTIE MATHÉMATIQUE, par M. D'ALEMBERT, de l'Académie Royale des Sciences de Paris, de celle de Pruſſe, & de la Société Royale de Londres.

Tantùm ſeries juncturaque pollet,
Tantùm de medio ſumptis accedit honoris! HORAT.

TOME PREMIER.

A PARIS,
Chez BRIASSON, *rue Saint Jacques, à la Science.*
DAVID l'aîné, *rue Saint Jacques, à la Plume d'or.*
LE BRETON, Imprimeur ordinaire du Roy, *rue de la Harpe.*
DURAND, *rue Saint Jacques, à Saint Landry, & au Griffon.*

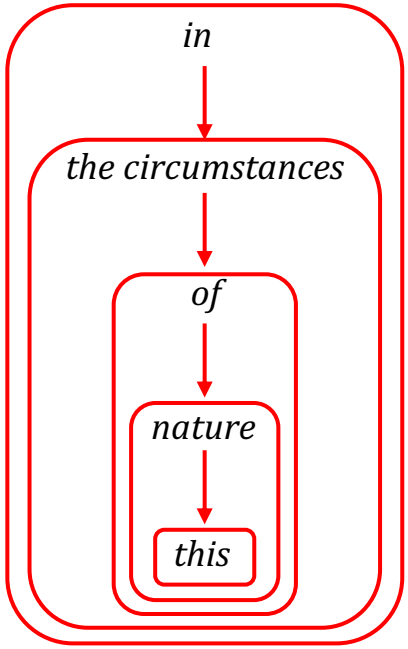
M. DCC. LI.
AVEC APPROBATION ET PRIVILEGE DU ROY.

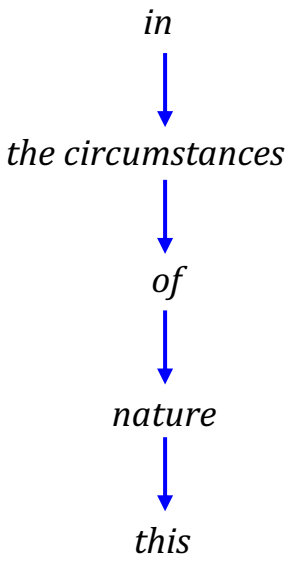
Beauzée 1765

- Nicolas Beauzée (article *Régime* (government) from the *Encyclopédie* of Diderot and D'Alembert, vol. 14, 1765)
  - "For instance in the sentence *with the care requested in the circumstances of this nature*; the word *nature* is the grammatical complement of the preposition *of*; *this nature* is its logical complement; the preposition *of* is the initial complement of the appellative noun *the circumstances*; and *of this nature* is its total complement; *the circumstances* is the grammatical complement of the preposition *in*; and *the circumstances of this nature* is its logical complement."

Beauzée 1765

"For instance in the sentence *with the care requested in the circumstances of this nature*; the word *nature* is the grammatical complement of the preposition *of*; *this nature* is its logical complement; the preposition *of* is the initial complement of the appellative noun *the circumstances*; and *of this nature* is its total complement; *the circumstances* is the grammatical complement of the preposition *in*; and *the circumstances of this nature* is its logical complement."





- grammatical complement = initial complement = dependent
- logical complement = total complement = constituent

Beauzée 1765

- Beauzée also gives what is probably the first definition of the **projectivity**:
  - “We never must break the unity of total complement by throwing another complement of the same word between its parts.”
    - I gave the book to a girl I met yesterday
    - *I gave to a girl, the book, I met yesterday
  - Today definition (Lecerf 1961): the projection of every word (= constituent it heads) is continuous
    - I gave [the book] [to a girl I met yesterday]
  - He adds that, contrary to rigid-order languages such as French, case-marking languages such as Latin can violate it.

Gaultier 1817

Grammar atlases or tables designed to stimulate and sustain children's attention in the study of grammar.

For *grammatical analysis*, you need a sheet of paper, a slate or a blackboard divided into ten columns. In the left-hand margin, write the words of the sentence to be analyzed, one below the other. In the first column, indicate to which of the three primary parts of speech each word belongs, and in the second to which of the ten secondary parts of speech each word belongs; in the third, fourth and fifth columns, indicate the gender, number and case of the nouns; in the sixth, seventh, eighth and ninth columns, indicate the number, person, tense and mode of the personal verb. In the tenth, all the divisions and subdivisions of the ten parts of speech are indicated. You can write only the initial letters of each word: subs. for substantive, etc. *Example:*

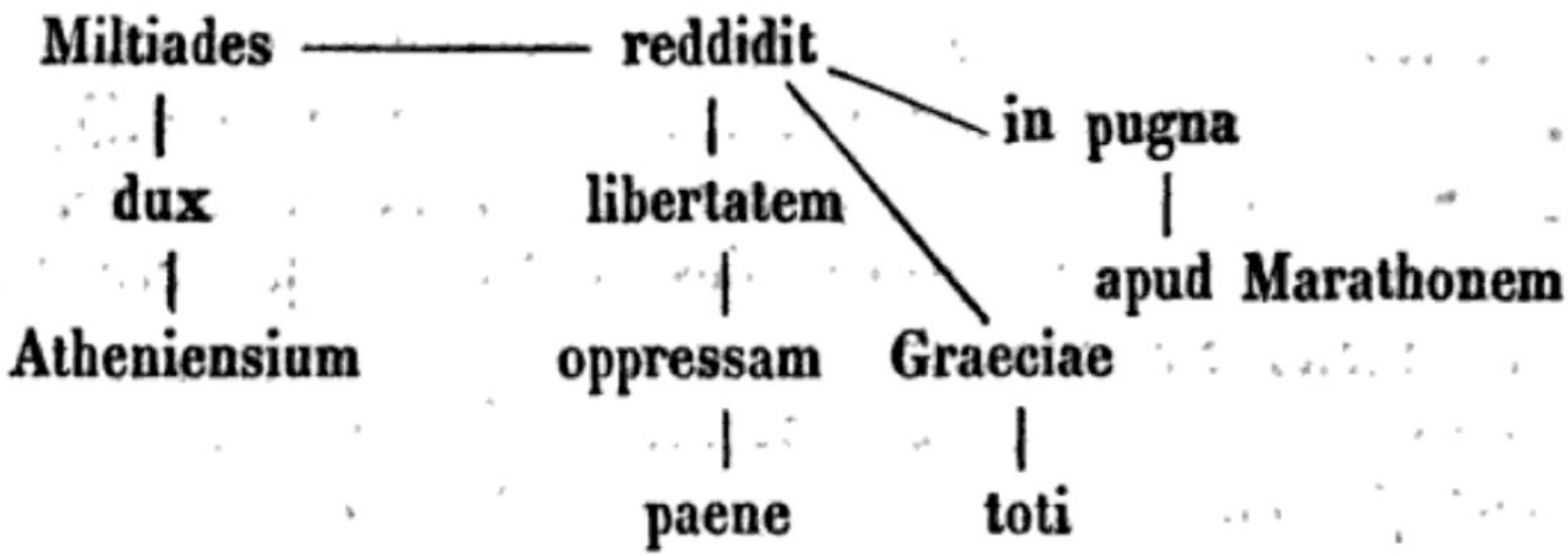
Gaultier 1817

	1	2	3	4	5	6	7	8	9	10
Paul	nom.	subst.	masc.	sing.	nom. de *vient*	. . .	. . .	. . .	. . .	pro.
ne	part.	adv.	. . .	. . .	. . .	. . .	. . .	. . .	. . .	nég.
vient	verb.	pers.	. . .	. . .	. . .	sing.	3e p.	pré.	ind.	2e c. neut.
plus	part.	adv.	. . .	. . .	. . .	. . .	. . .	. . .	. . .	nég.
te	nom.	pro.	masc.	sing.	acc. de *voir*	. . .	. . .	. . .	. . .	pers.
voir.	verb.	inf.	. . .	. . .	. . .	. . .	. . .	. . .	. . .	3e c. act.

Paul ne vient plus le voir
Paul not comes anymore him see
‘Paul doesn’t come anymore to see him’

Gustav Billroth (1832)

- First known syntactic diagram
  - Miltiades, dux Atheniensium, toti Graecia libertatem paene oppressam in pugna apud Marathonem reddidit (*Lateinishe Schulgrammatik*, p. 329)



Barnard 1836

- Frederik A. Barnard, 1836, *Analytic Grammar with Symbolic Illustrations*
  - English professor for deaf people

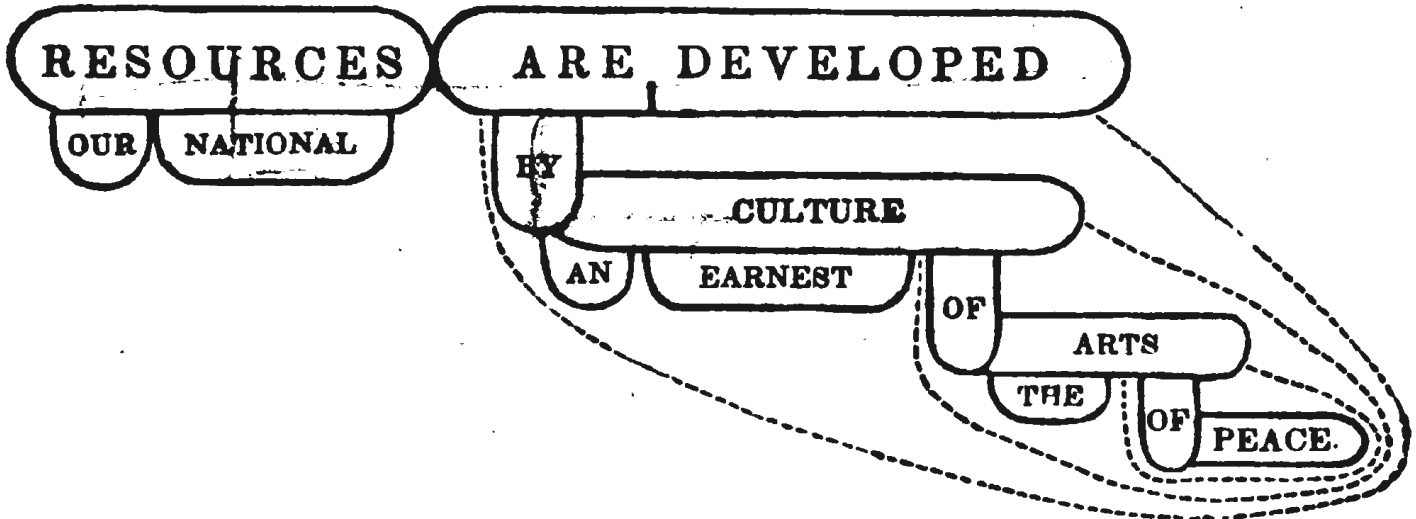
Clark 1847

1. *" The king of shadows loves a shining mark."*

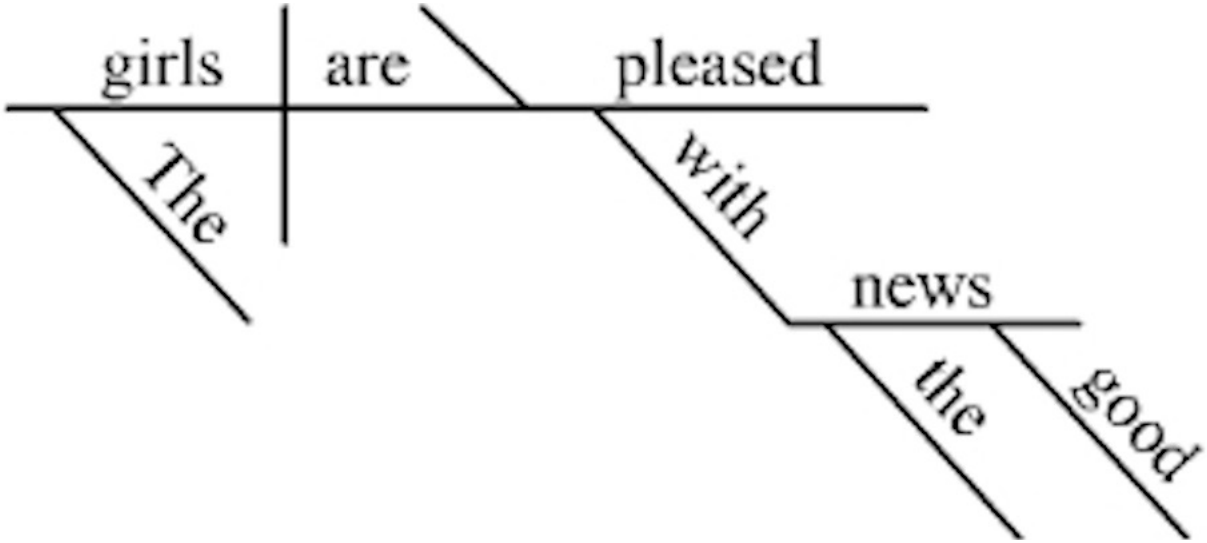
(13.)

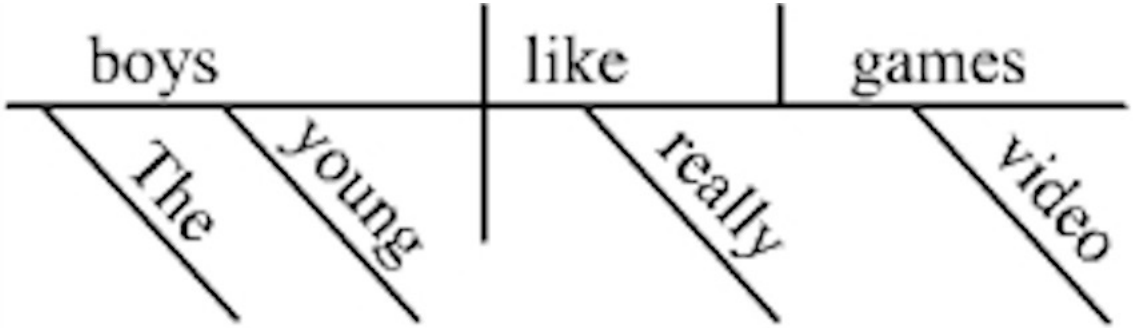
" Our national resources are developed by an earnest culture of the arts of peace."

(4.)



Reed & Kellogg 1877

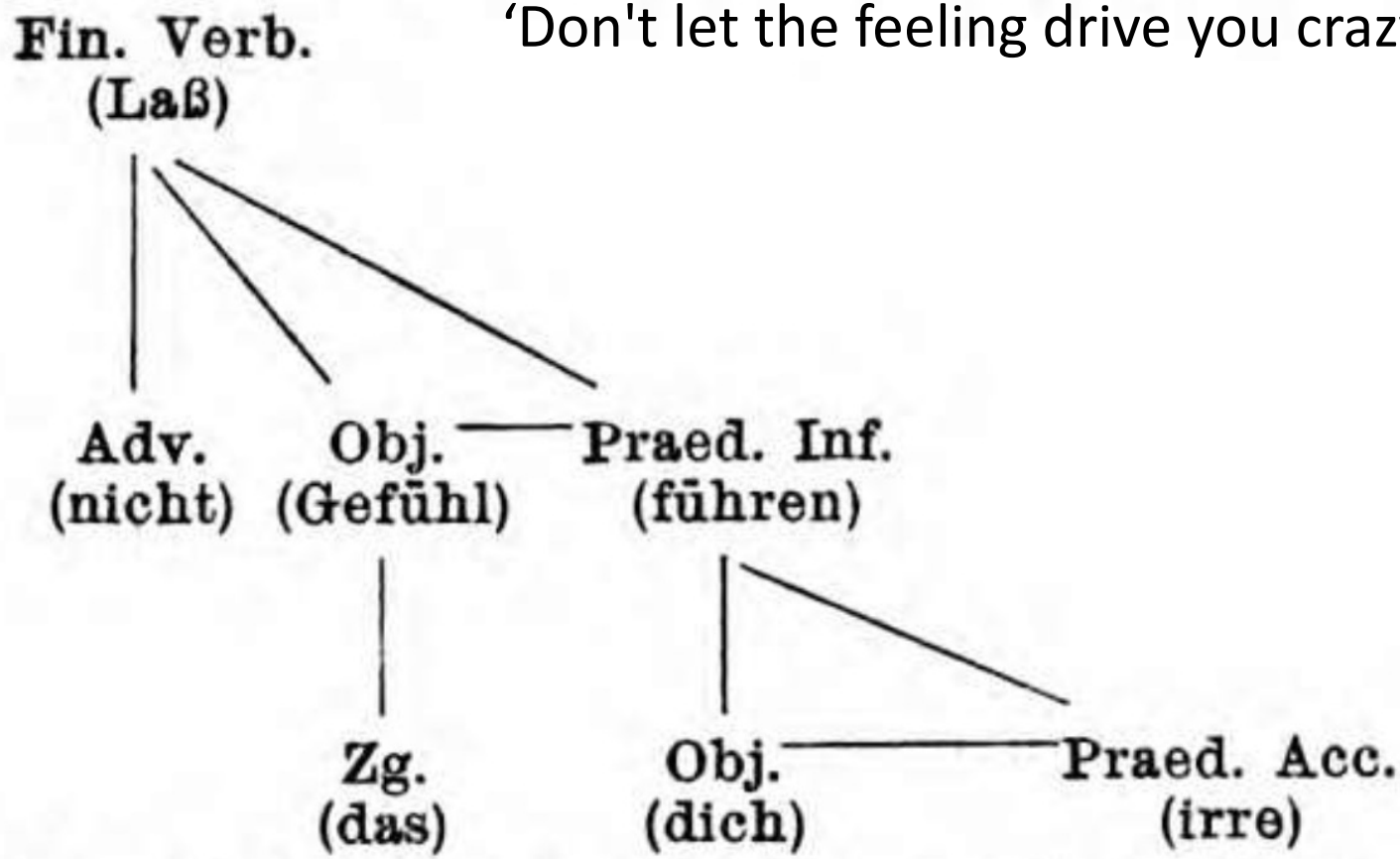


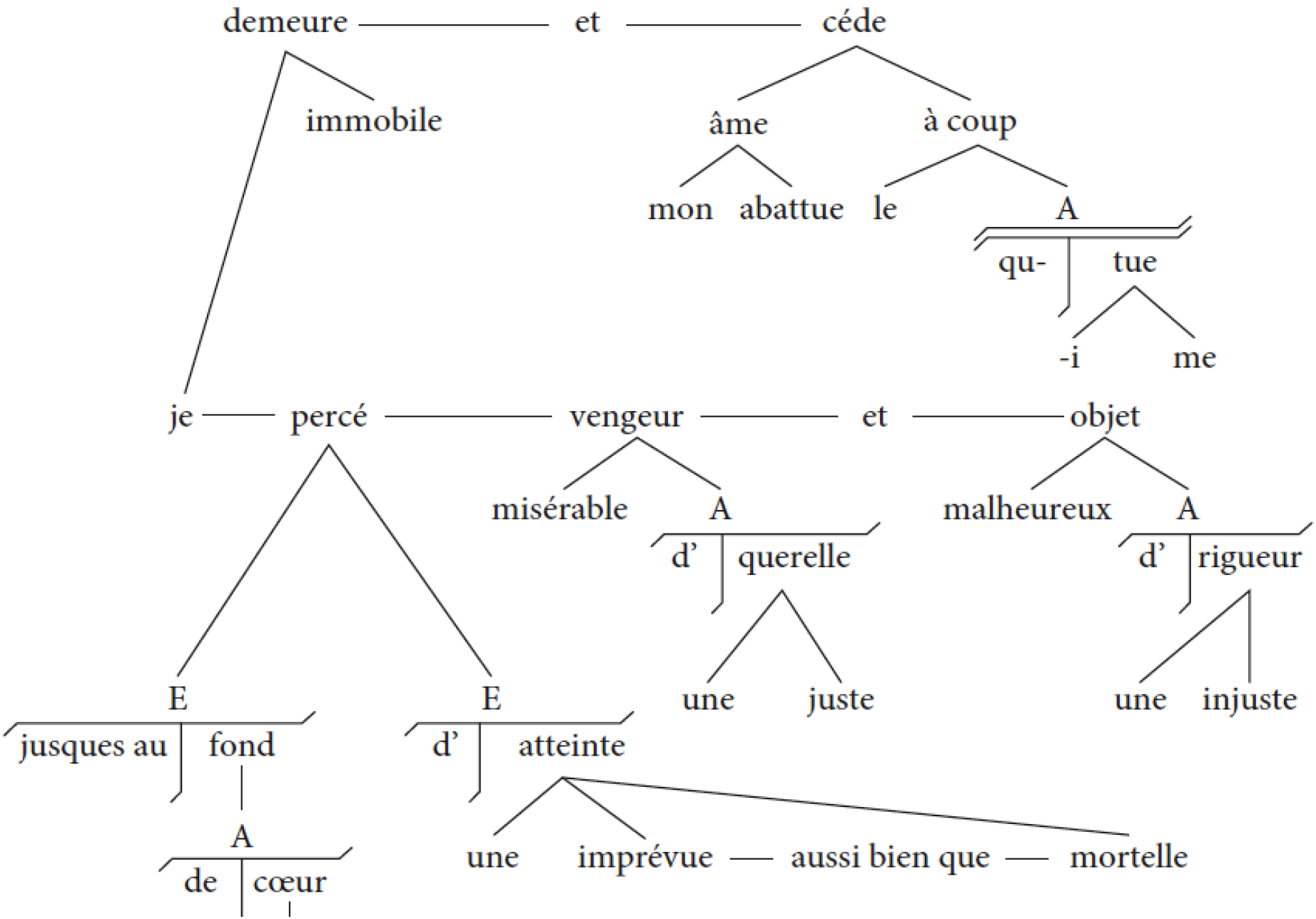


Kern 1883

„Laß nicht das Gefühl dich irre führen.“

‘Don't let the feeling drive you crazy’





Tesnière 1959

Corneille, *Le Cid, I, 6*, Stemma 361

Treebanks

- a syntactic treebank is
  - a corpus
  - fully annotated with syntactic structures

Treebanks

8: avec sucre et sans sucre gold x

9: merci gold x

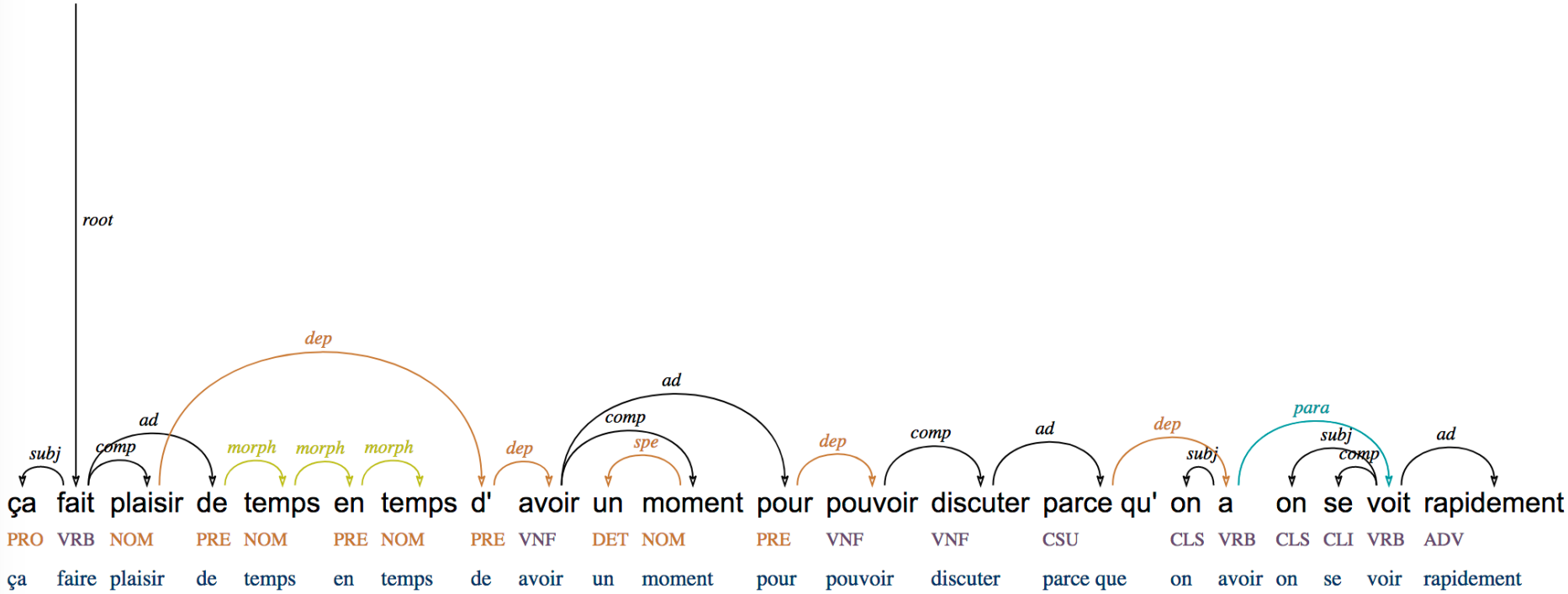
10: sucre ouais gold x

11: oui merci merci gold x

12: sans sucre toi gold x

13: voilà tu vois donc écoute gold x

14: ça fait plaisir de temps en temps d' avoir un moment pour pouvoir discuter parce qu' on a on se voit rapidement gold x



15: bé c' est vrai que en plus depuis depuis septembre on a pas eu l' occasion de se voir hein gold x

16: donc gold x

17: devant les bureaux gold x

18: des trucs rapides gold x

Jespersen (1937, *Analytic syntax)*

8. 8. O/S

What ails Tom? S/O? V O/S.

(The old "What aileth thee?" has become „What ailst thou?")

G. Mich friert; mir graut O/S V.

G. Mich jammert seiner O/S V 3/O.

L. Pudet eum sceleris V O/S 3/O.

G. Mich reut dieser tat O/S V 3/O(21).

F. Me voici O/S {V3}.

It. Eccolo {3 O/S}.

It. Questo non si dice S/O 3ⁿ O/S V.

It. Si vende carne; Si vendono biglietti O/S V S/O.

By a shifting (see PhilGr 161) this leads to Si viene S V; Si vende biglietti S V O, in which *si* must be considered S with the same generic signification as F. *on*. Cf. above on F. Cela ne se dit pas.

Sp. Se le trató como á un rey 'on l'a traité comme un roi'

S/O O V 3ᶜ pl.

Sp. Se conoce al verdadero amigo en la necesidad S/O V pO(21) pl.

Dan. Mig synes du har ret (= G. Mir scheint, du hast recht)

O V S(S_2 V O_2) has become: Jeg synes du har ret

S V O (S_2 V O_2).

Digital treebanks

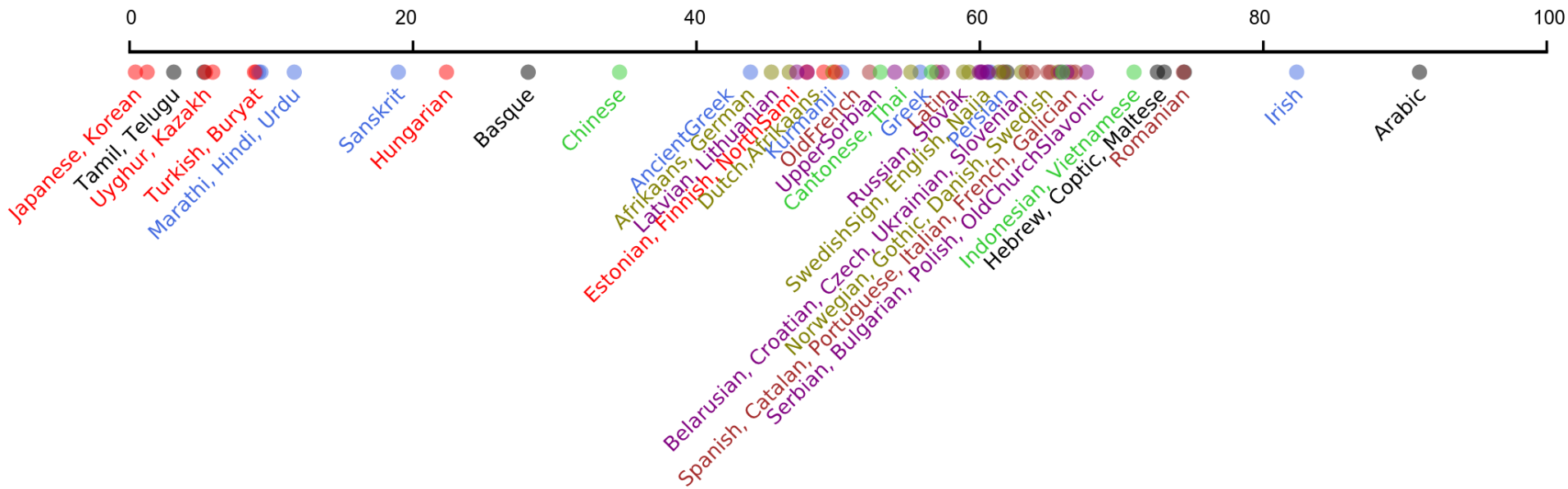
1970s : Talbanken (Swedish)
1989-1996: Penn Tree Bank (English)
1997: Negra Treebank (German)
1995-now: Prague Dependency Treebank (Czech)
2003: French Tree Bank (French)
~ 2005: Dependency parsing becomes dominant
2005: the Stanford parser (2002) proposes a dependency-based output
2007: CoNLL dataset => CoNLL format for dependency trees
2008: POS interset, many projects of conversion
2014: Google provides treebanks for 30 languages (based on Stanford schema)
2014: UD starts

What are treebanks for?

- computational parsing
  - treebank => parser
  - parser => treebank
- Linguistic studies
  - syntax, intono-syntax, syntax-semantics interface
  - language typology
  - (quantitative) grammars
  - psycho-linguistics

Treebanks and typology

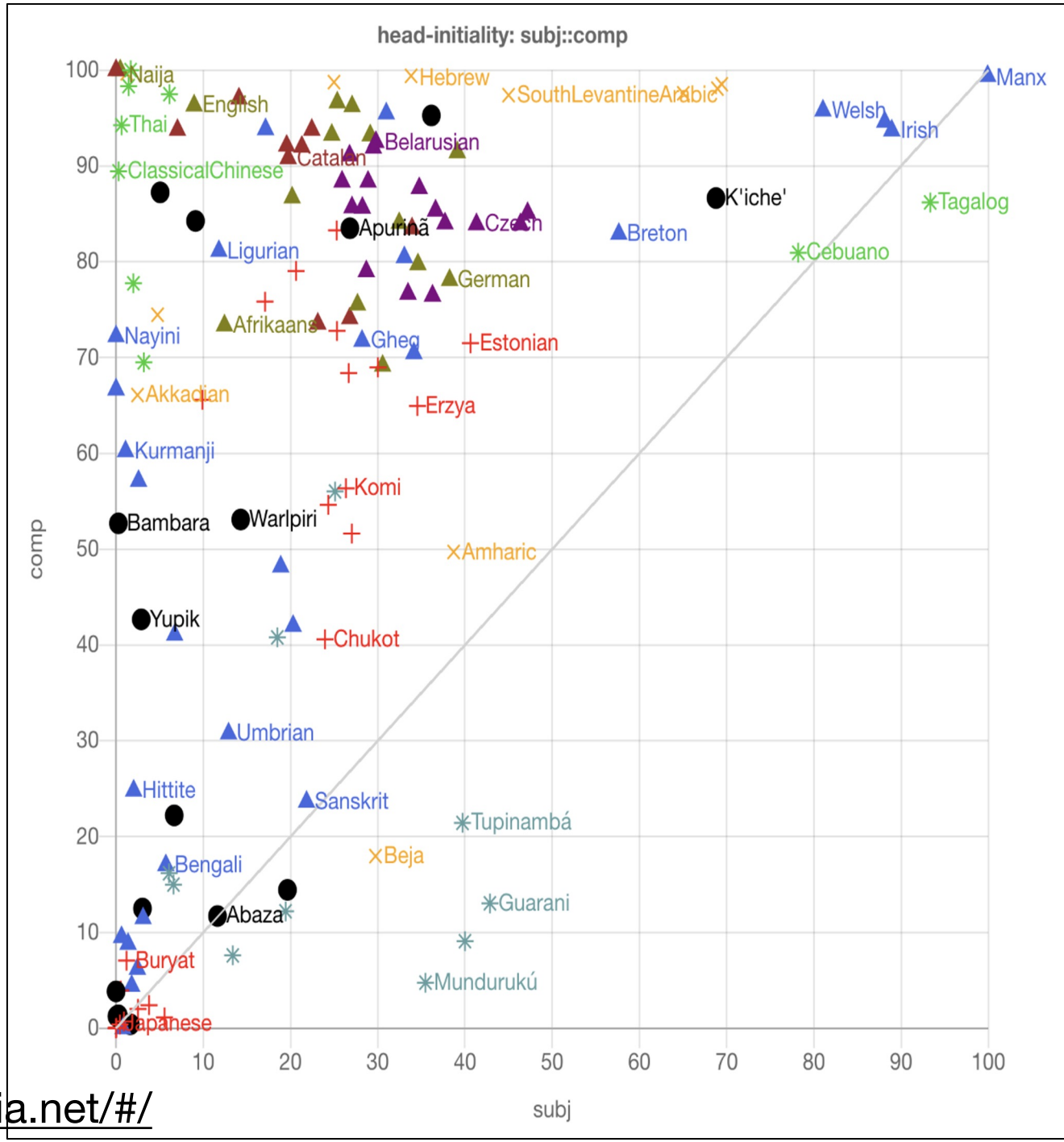
Typology



%-age of dependents on the right of their governor

Gerdes, Kahane & Chen (2021) Typometrics: From Implicational to Quantitative Universals in Word Order Typology, *Glossa: a journal of general linguistics 6(*1): 16. 1–31. DOI: http://doi.org/10.5334/gjgl.764

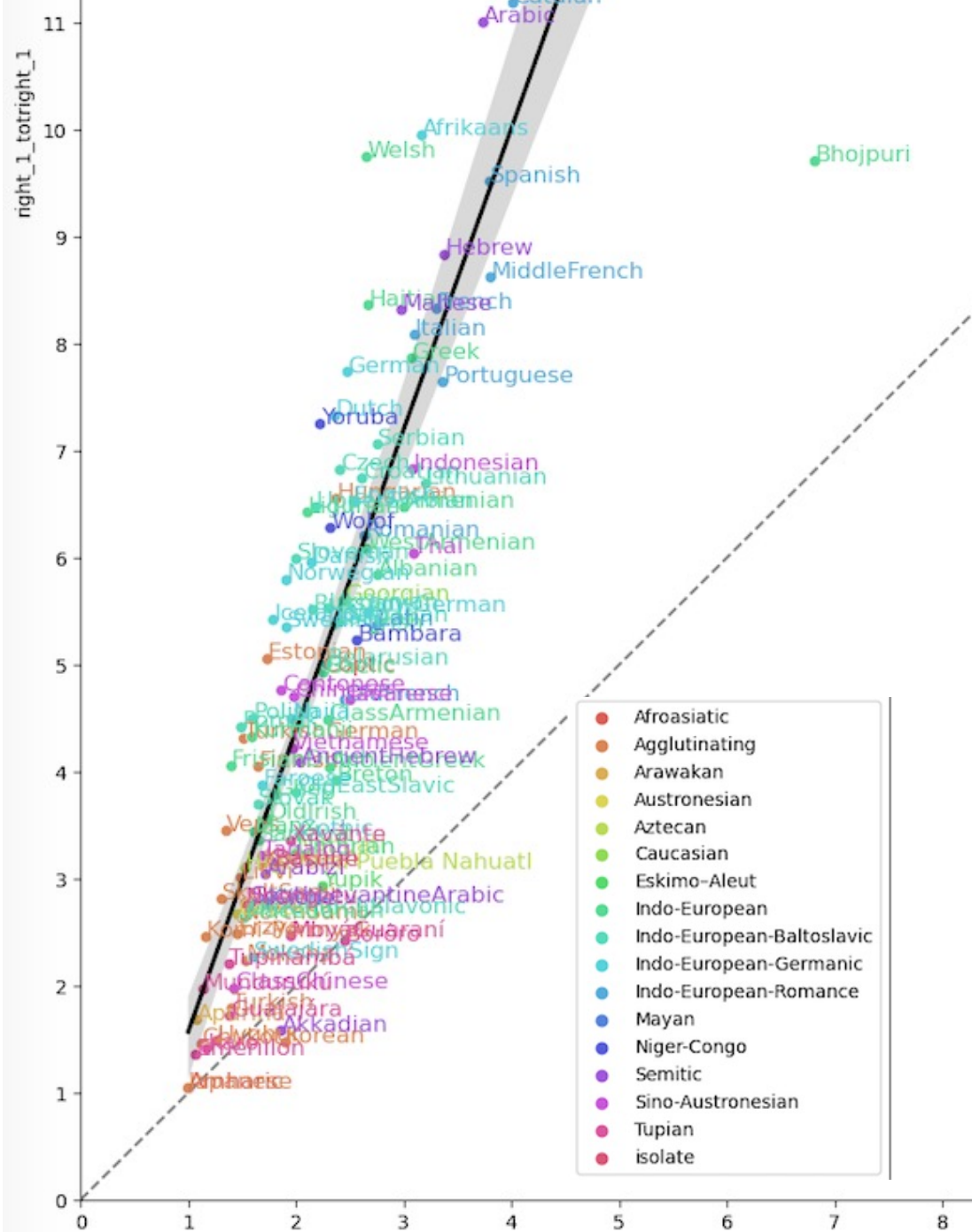
Percentage of subjects and complements on the right of the verb



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

length of the first constituent on the right of the verb
X = with a second constituant after it
Y = without another constituent

Chen X., Gerdes K., Kahane S., Courtin M. (2021) [The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages]. *Proceedings of Qualico*, 15 p.



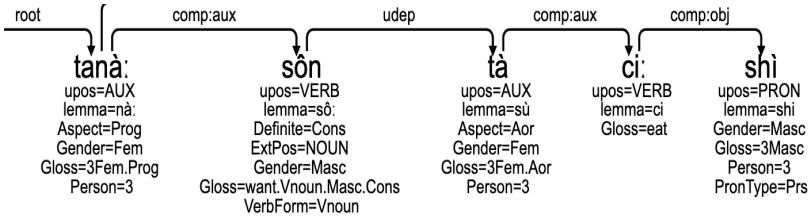
Treebanks and grammars

Induction of grammars

- Charniak 1996
  - treebank => PCFG => parser
  - PCFG = Probabilistic Context-Free Grammar
    = CFG with a weight (frequency) on each rule
- similar works with TAG, LFG, HPSG …
- but such grammars are parsing-oriented
  - grammar = a bag containing numerous formal rules
  - not easy to understand what are the main properties
- what about descriptive grammars (i.e. human-oriented grammars)?
  - we want to know what the main properties of a given language are

Autogramm projet anr (2022-2026)

- Induction of descriptive grammars from syntactic treebanks

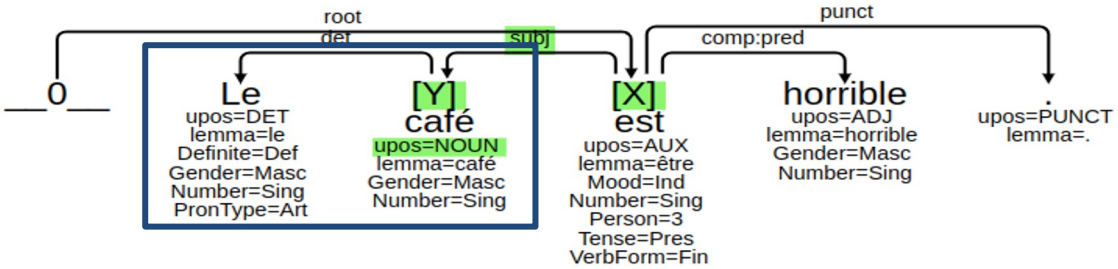


extraction of grammatical description

Most salient patterns answering the question

Example of a (quantitative) descriptive grammar

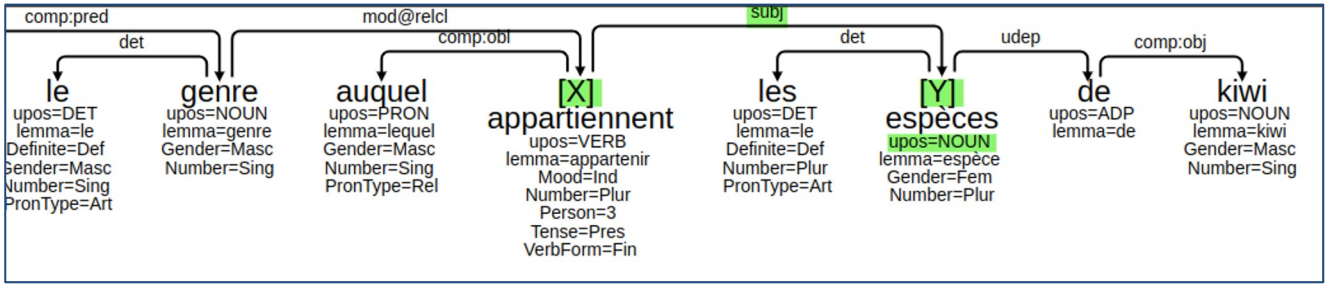
- French is SVO



97% of subjects are before the verb

3% are inverted
Question: When do we do that?

Possible answer: 23% of nominal subjects in relative clauses are inverted



Contrastive grammars

- What are the rules that distinguish French and English?
- Example: What are the differences in the verb construction?
  - Put together a treebank of French and a treebank of English
  - Question : given a verb, how I know I am in the French treebank?
  - Possible answer: I have pronouns before the verb
    - *Je lui parle* ‘I talk to her’, lit. I to_her talk

Contrastive grammars

- What are the rules that distinguish French and English?
- Example: What are the differences in the verb construction?
  - Put together a treebank of French and a treebank of English
  - Question : given a verb, how I know I am in the English treebank?
  - Possible answer: I have adverbs before the verb
    - *I often do that* vs *Je fais souvent ça*, lit. I do often that

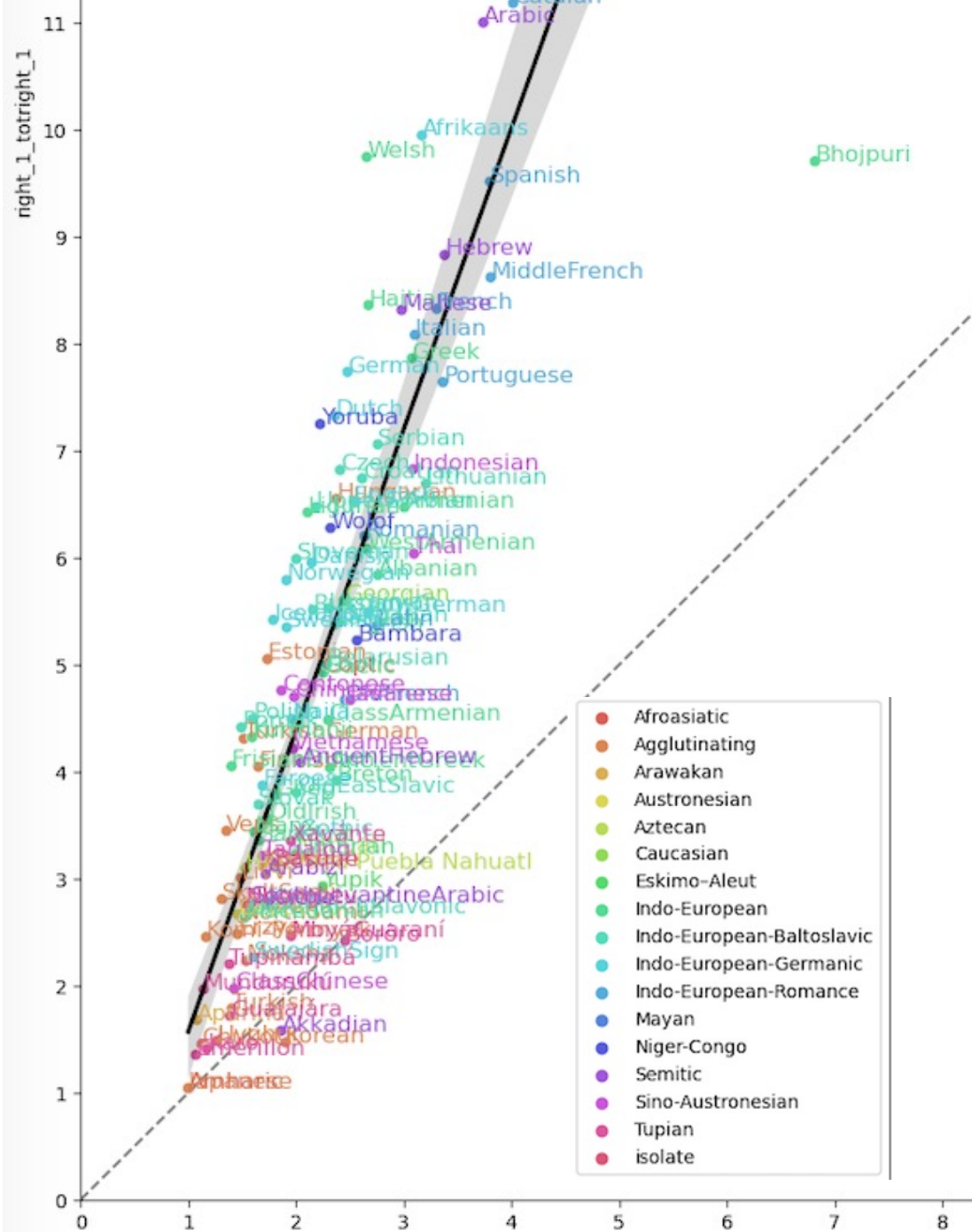
Treebanks and psycholinguistics

Dependency length minimization (DLM)

- Dependency lengths tend to be minimized in linguistic productions (Hudson 1984, Gibson 1998, Liu 2008 ; Futrell et al. 2015, 2020)
- **Properties correlated with DLM**
- Heavy constituent shift: when there are two constituents after a verb, the second tend to be heavier than the first
- Much less non-projective structures in natural languages than in randomly ordered trees (Ferrer i Cancho, 2006 ; Liu, 2008)
- DLM is a factor affecting the grammar of languages and word order choices (Gildea & Temperley, 2010 ; Temperley & Gildea, 2018)

length of the first constituent on the right of the verb
X = with a second constituant after it
Y = without another constituent

Chen X., Gerdes K., Kahane S., Courtin M. (2021) [The Co-Effect of Menzerath-Altmann Law and Heavy Constituent Shift in Natural Languages]. *Proceedings of Qualico*, 15 p.



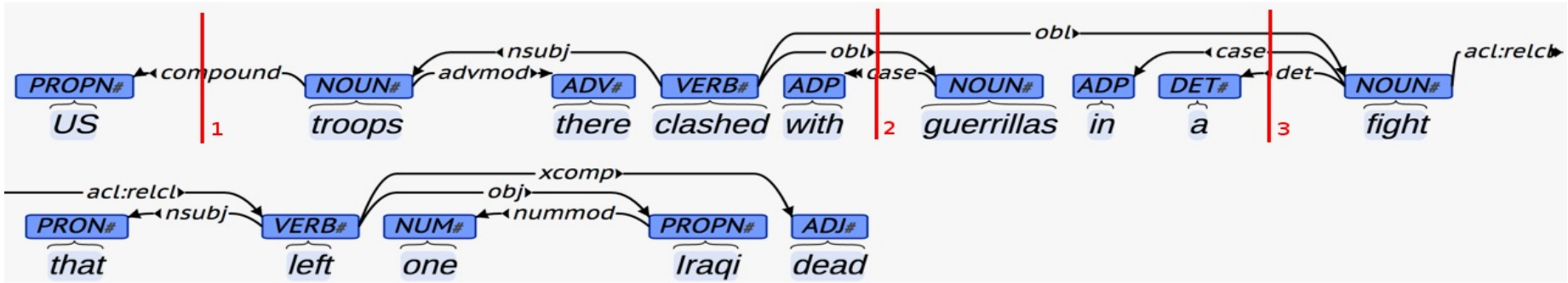
DLM and psycholinguistics

- the longer the dependency is, the longer the information must be kept in the memory
  - short-term memory

DLM and dependency flux

dependency flux between two words = set of dependencies that link a word on the left with a word on the right (Kahane et al., 2017).

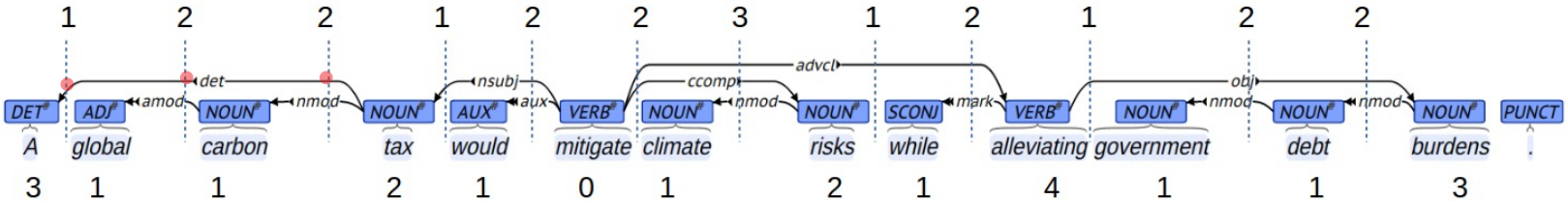
flux size at position P = number of dependencies that cross P



Position 1: flux size = 1 Position 2: flux size = 3 Position 3: flux size =3

DLM and dependency flux

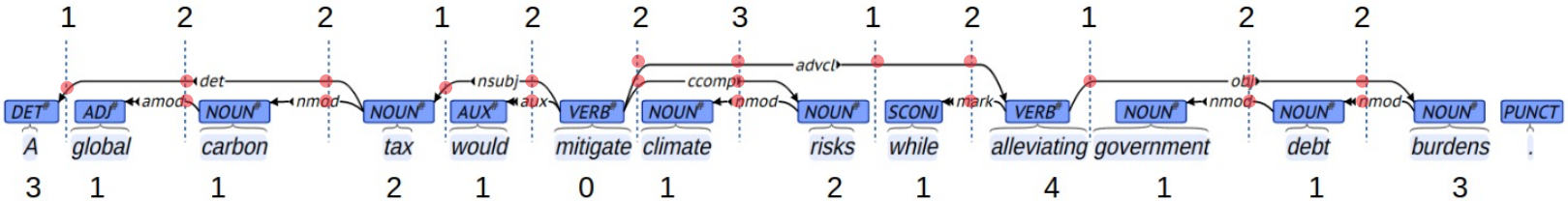
- Average dependency length = average dependency flux



- **relation *det***
  - **length = 3**
  - **= cross 3 inter-word fluxes (red points)**

DLM and dependency flux

- Average dependency length = average dependency flux



- sum of red points = sum of dependency lengths
  = sum of flux size

(Kahane, Yan 2017,2019 ; Yan 2021)

DLM and psycholinguistics

- the longer the dependency is, the longer the information must be kept in our memory
  - short-term memory
- the larger the dependency flux is, the more information we need to keep in memory
  - the size of the short-term memory is very small (Miller 1956, The magical number seven, plus or minus two: Some limits on our capacity for processing information)

Conclusion

- treebanks for NLP
  - treebank => parser => treebank
- treebanks for linguistics
  - quantitative typology
  - quantitative grammar
  - contrastive grammar
- treebanks for psycholinguistics
  - dependency length minimization and flux size minimization