TOKEN-BASED TYPOLOGY IN PRACTICE: A WORKSHOP WITH MINICIEP+

Lecturer: Luigi Talamo (UdS, Germany)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LABORATORIO
SPERIMENTALE

UNIVERSITÄT
DES
SAARLANDES

OUTLINE

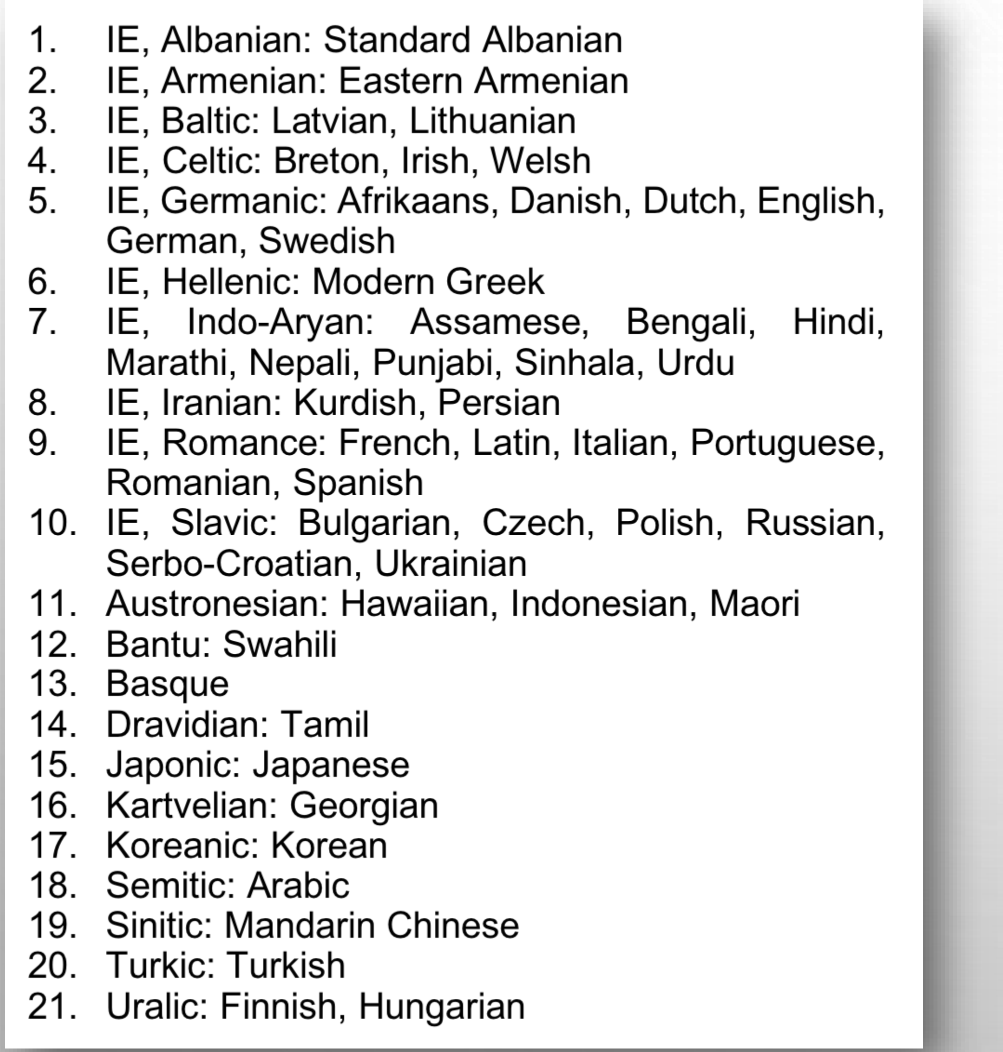
- **What is miniCIEP+**
- **A very short introduction to Universal Dependencies (UD)**
- **Experimenting with miniCIEP+**

WHAT IS MINICIEP?

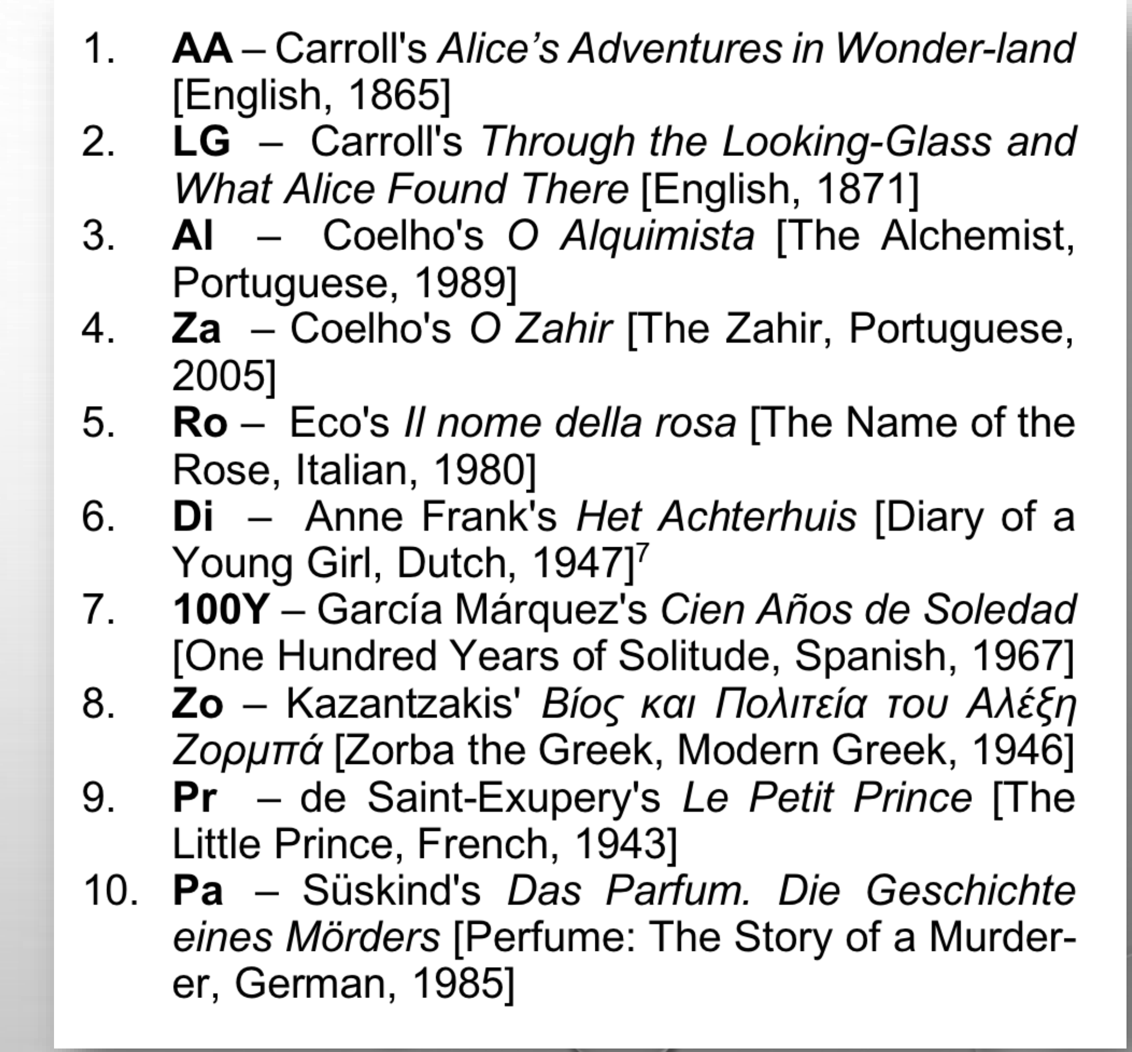
A SHARABLE PARALLEL CORPUS OF PROSE

- Started at Saarland University in Autumn 2019 – Credits: Annemarie Verkerk (PI), Luigi Talamo (Post Doc) and Andrew Dyer (PhD candidate)
- Derivative of CIEP+; the Corpus of Indo-European Prose Plus /kiːp plʌs/
- Contents: contains about 14% of 10 frequently translated literary works
- Language sample: 35 Indo-European; 15 non-IE languages
- Size: subcorpora typically ~ 5750 sentences and up to 125K tokens
- Annotation in the Universal Dependencies format + information status
- Sharable: we offer considerations of German law as to what constitutes “a select group of people”
- Status: mini-CIEP+ v. 1.0 contains 35 languages

WHAT’S INSIDE



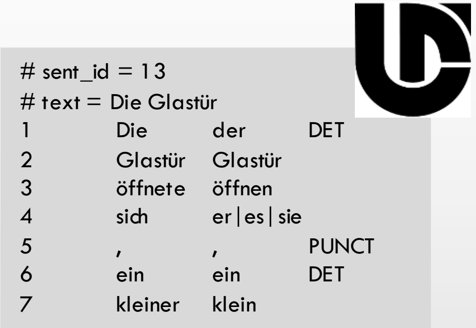
1. IE, Albanian: Standard Albanian
2. IE, Armenian: Eastern Armenian
3. IE, Baltic: Latvian, Lithuanian
4. IE, Celtic: Breton, Irish, Welsh
5. IE, Germanic: Afrikaans, Danish, Dutch, English, German, Swedish
6. IE, Hellenic: Modern Greek
7. IE, Indo-Aryan: Assamese, Bengali, Hindi, Marathi, Nepali, Punjabi, Sinhala, Urdu
8. IE, Iranian: Kurdish, Persian
9. IE, Romance: French, Latin, Italian, Portuguese, Romanian, Spanish
10. IE, Slavic: Bulgarian, Czech, Polish, Russian, Serbo-Croatian, Ukrainian
11. Austronesian: Hawaiian, Indonesian, Maori
12. Bantu: Swahili
13. Basque
14. Dravidian: Tamil
15. Japonic: Japanese
16. Kartvelian: Georgian
17. Koreanic: Korean
18. Semitic: Arabic
19. Sinitic: Mandarin Chinese
20. Turkic: Turkish
21. Uralic: Finnish, Hungarian



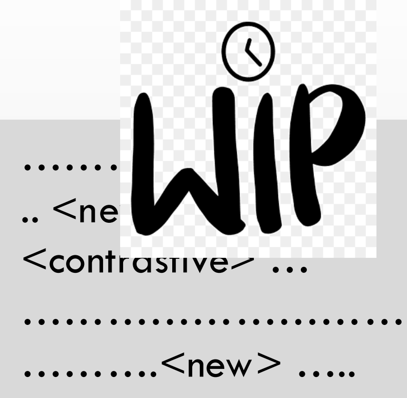
1. **AA** – Carroll's *Alice’s Adventures in Wonder-land* [English, 1865]
2. **LG** – Carroll's *Through the Looking-Glass and What Alice Found There* [English, 1871]
3. **Al** – Coelho's *O Alquimista* [The Alchemist, Portuguese, 1989]
4. **Za** – Coelho's *O Zahir* [The Zahir, Portuguese, 2005]
5. **Ro** – Eco's *Il nome della rosa* [The Name of the Rose, Italian, 1980]
6. **Di** – Anne Frank's *Het Achterhuis* [Diary of a Young Girl, Dutch, 1947][7]
7. **100Y** – García Márquez's *Cien Años de Soledad* [One Hundred Years of Solitude, Spanish, 1967]
8. **Zo** – Kazantzakis' *Βίος και Πολιτεία του Αλέξη Ζορμπά* [Zorba the Greek, Modern Greek, 1946]
9. **Pr** – de Saint-Exupery's *Le Petit Prince* [The Little Prince, French, 1943]
10. **Pa** – Süskind's *Das Parfum. Die Geschichte eines Mörders* [Perfume: The Story of a Murderer, German, 1985]

WHAT YOU GET (AND HOW WE DID IT)

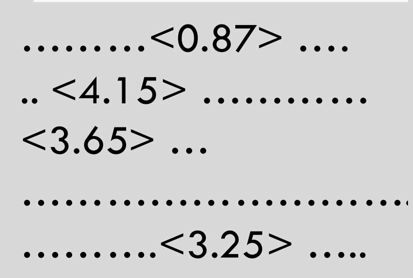
Multi-layer and modular structure



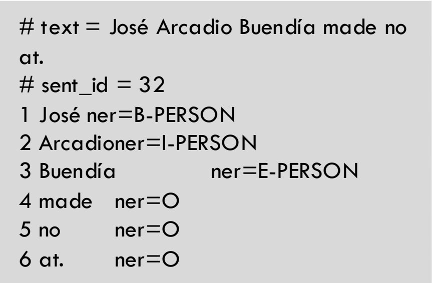
```
# sent_id = 13
# text = Die Glastür
1    Die       der        DET
2    Glastür   Glastür
3    öffnete   öffnen
4    sich      er|es|sie
5    ,         ,          PUNCT
6    ein       ein        DET
7    kleiner   klein
```



```
…….
.. <ne
<contrastive> …
…………………………
……….<new> …..
```



```
………<0.87> ….
.. <4.15> ………..
<3.65> …
…………………………
……….<3.25> …..
```



```
# text = José Arcadio Buendía made no
at.
# sent_id = 32
1 José ner=B-PERSON
2 Arcadioner=I-PERSON
3 Buendía          ner=E-PERSON
4 made    ner=O
5 no      ner=O
6 at.     ner=O
```

Metadata

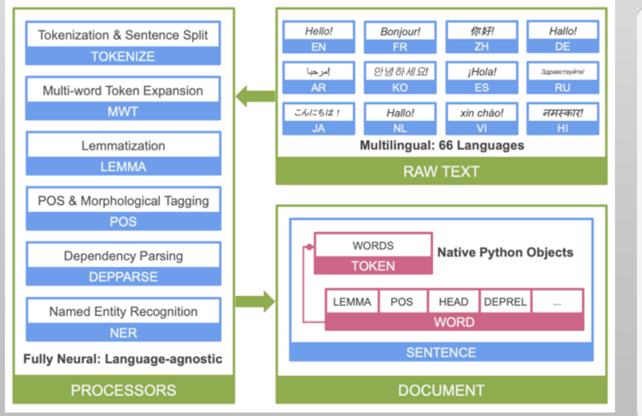
Universal Dependencies

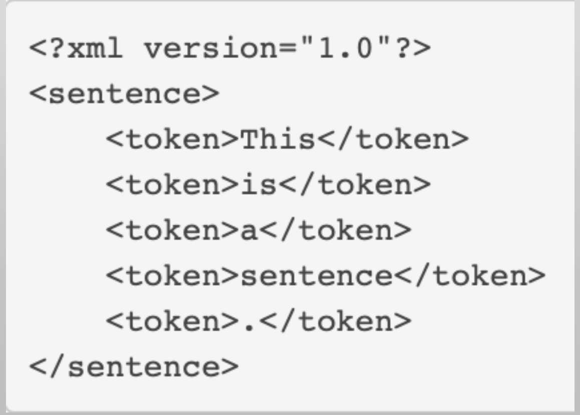
information status

surprisal

Named Entity Recognition

Tools





```
<?xml version="1.0"?>
<sentence>
    <token>This</token>
    <token>is</token>
    <token>a</token>
    <token>sentence</token>
    <token>.</token>
</sentence>
```

CQPweb; hosted by Prof. Teich

Python: converting between formats; Python Stanza library for UD parsing (pyconll, conllu)

xml for other annotation layers, CQPweb for querying some of the relevant layers together

HOW WE DID IT

…steps in creating CIEP+ and mini-CIEP+

1. obtain a physical copy of each book (the university library now owns some antiques, rarities, illustrated works… some come with great stories)
2. create or buy in addition a digital version of each book; in most cases this means OCR + OCR correction by a human annotator
3. add metadata and catalogue the physical books in the university library
4. use the Stanford Stanza natural language analysis package to parse the texts (sentence splitting, tokenization, lemmatization, parts-of-speech and syntactic dependencies tagging)
5. find solutions for sampled languages without a pretrained Stanza parser and/or without a UD treebank (creating treebanks ourselves)

HOW CAN WE SHARE MINICIEP+ WITH YOU?

- Hartmann (2023): “The replication crisis in linguistics is highly relevant to corpus-based research: Many corpus studies are not directly replicable as the data on which they are based are not readily available.”
- German copy-right law (Urheberrecht)§ 60c and 60d: “*For the purpose of non-commercial scientific research, up to 15 percent of a work may be reproduced, distributed and made publicly accessible […] to a defined circle of people for their own scientific research*”
- Audience mini-CIEP+: corpus-based typologists, contrastive linguists and language specialists, especially for low-resourced languages;
- Condition: data usage agreement that specifies exactly what the researchers need; and how they are supposed to make sure it does not become public.

A VERY SHORT INTRODUCTION TO UNIVERSAL DEPENDENCIES (UD)

UNIVERSAL DEPENDENCIES

- Why a dependency treebank? Pros and cons according to Daniel Zeman (https://ufal.mff.cuni.cz/~zeman/2023/docs/1-introduction.pdf)
  - Economical, free word order, head of a phrase ✅
  - No derivation history, coordination/apposition, secondary predicates (two dependencies) ❌
- But, most important, why Universal Dependencies?
  - ‘universal’, lots of languages (over 150 languages);
  - widely employed (over 200 treebanks);
  - several layers of annotation.

UNIVERSAL DEPENDENCIES

de Marneffe, Marie-Catherine; Manning, Christopher D.; Nivre, Joakim & Zeman, Daniel 2021. Universal dependencies. Computational Linguistics 47,2. 255-308. From the abstract:

"Universal dependencies (UD) is a framework for **morphosyntactic annotation** of human language, which to date has been used to create treebanks for more than 100 languages. In this article, we outline the linguistic theory of the UD framework, which draws on a long tradition of **typologically oriented grammatical** theories. Grammatical relations **between words** are centrally used to explain how **predicate–argument structures are encoded morphosyntactically** in different languages while **morphological features and part-of-speech classes** give the properties of words. We argue that this theory is a good basis for crosslinguistically consistent annotation of typologically diverse languages in a way that supports **computational natural language understanding** as well as **broader linguistic studies.**”

UNIVERSAL DEPENDENCIES: BASIC TENETS

- **Dependency grammar: head** and **dependent;**
- Three fundamental units: **nominal (entity)**, **clause (event)** and **modifier (attribute);**
- **Words** (tokens) **as basic units;**
- **Grammatical relations** are between **words.**

UNIVERSAL DEPENDENCIES: BASIC TENETS

Head and Dependents

Binary grammatical relation: an arrow goes from the head to the dependent and is labelled for a grammatical relation.

How do we identify the head?

Nominal phrases: noun;

Clause: usually **verbs,** but could be also **nominals** or **adjectives.**

When in doubt, the element with most important **content/meaning** is the **head.**

UNIVERSAL DEPENDENCIES: BASIC TENETS

Head and Dependents

Binary grammatical relation: an arrow goes from the head to the dependent and is labelled for a grammatical relation.

How do we identify the head?

Nominal phrases: noun; *The good* ***doctor***

Clause: usually **verbs,** but could be also **nominals** or **adjectives.** *The good doctor* ***visits*** *her patients*

Adjectives: *The doctor is* ***good.*** **Nominals:** *My sister is a good* ***doctor.***

When in doubt, the element with most important **content/meaning** is the **head.**

The good doctor has ***arrived.***

UNIVERSAL DEPENDENCIES: BASIC TENETS

Nominals, clause and modifiers

Nominals: default/canonical items for referring to an entity

Clause: default/canonical items for referring to event

Modifiers: default/canonical items for modifying a clause, a nominal or another modifier

UNIVERSAL DEPENDENCIES: BASIC TENETS

Nominals, clause and modifiers

Nominals: default items for referring to an entity **Reference**

Clause: default items for referring to event **Predication**

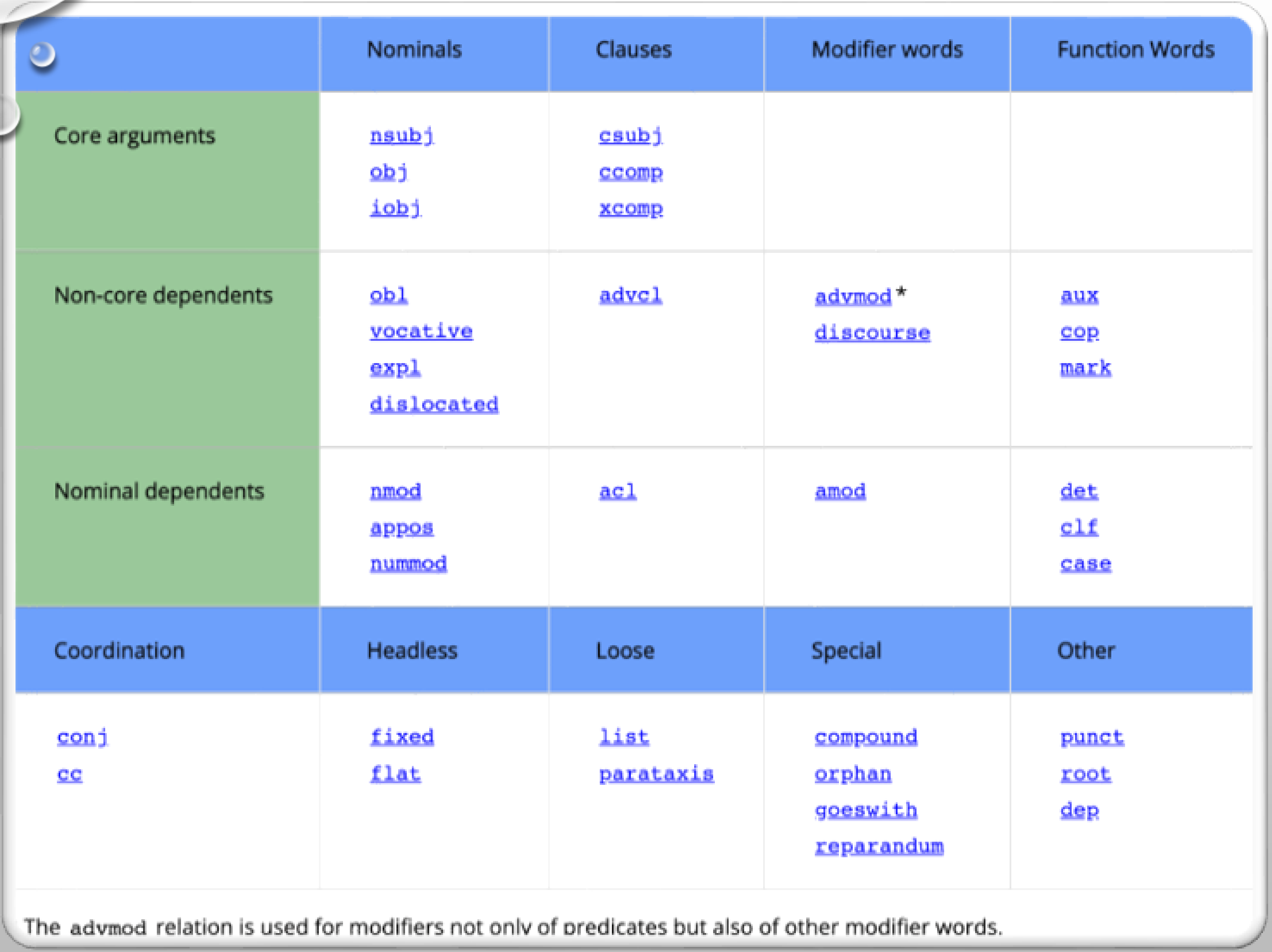
Modifiers: default items for modifying a clause, a nominal or another modifier **Modification**

This may remind some of you of **Croft's propositional acts** / **information packaging functions!**

UNIVERSAL DEPENDENCIES: CONLL-U FILES

Ten fields for the annotation, separated by single tab characters:

1. ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes (decimal numbers can be lower than 1 but must be greater than 0).
2. FORM: Word form or punctuation symbol.
3. LEMMA: Lemma or stem of word form.
4. **UPOS: Universal part-of-speech tag.**
5. XPOS: Optional language-specific (or treebank-specific) part-of-speech / morphological tag; underscore if not available.
6. **FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.**
7. **HEAD: Head of the current word, which is either a value of ID or zero (0).**
8. **DEPREL: Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one.**
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
10. MISC: Any other annotation.



	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case

Coordination	Headless	Loose	Special	Other
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

The advmod relation is used for modifiers not only of predicates but also of other modifier words.

UNIVERSAL DEPENDENCIES: BASIC TENETS

There is a fundamental distinction between Nominals and Clauses

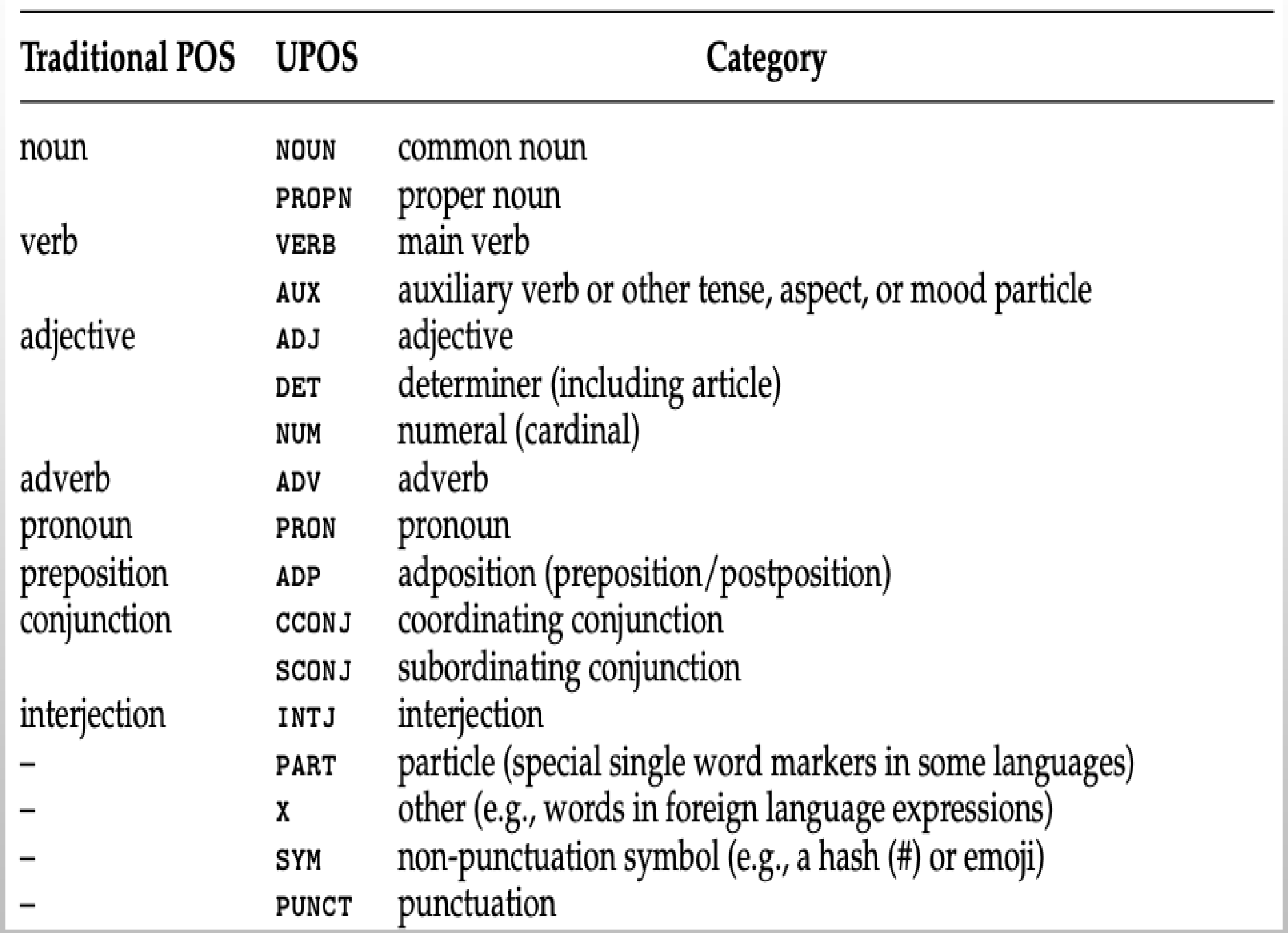
UNIVERSAL DEPENDENCIES: UPOS

Universal Parts of Speech (UPOS)

- Words can be classified into categories: **lexical categories** aka **word categories** aka **parts of speech.**
- These categories are not universal **but language-specific.** Still, if we want to use the same set of categories, *we have to live with that*. There are **17 UPOSes** in UD, defining both words and elements of text such as punctuations or symbols.
- We fit **language-specific categories** into these universal categories using several approaches:
  - a semantic approach: **nouns** usually -> **objects, verbs -> actions** and **adjectives -> properties.**
  - A **distributional** approach:
    - **Syntactic** and **morphological** properties: i.e., nouns usually pop up as **verbal arguments,** they inflect for given features in the language X, …

UNIVERSAL DEPENDENCIES: UPOS

Universal Parts of speech (UPOS)



Traditional POS	UPOS	Category
noun	NOUN	common noun
	PROPN	proper noun
verb	VERB	main verb
	AUX	auxiliary verb or other tense, aspect, or mood particle
adjective	ADJ	adjective
	DET	determiner (including article)
	NUM	numeral (cardinal)
adverb	ADV	adverb
pronoun	PRON	pronoun
preposition	ADP	adposition (preposition/postposition)
conjunction	CCONJ	coordinating conjunction
	SCONJ	subordinating conjunction
interjection	INTJ	interjection
–	PART	particle (special single word markers in some languages)
–	X	other (e.g., words in foreign language expressions)
–	SYM	non-punctuation symbol (e.g., a hash (#) or emoji)
–	PUNCT	punctuation

https://universaldependencies.org/u/pos/all.html

UNIVERSAL DEPENDENCIES: MORPHOLOGICAL FEATURES (FEATS)

Universal morphological features

- As the name suggests, this annotation field concerns the **features of the word**: **nominal, adjectival and verbal categories** such as gender, degree and tense.
- TBH, this is a bit of misnomer, as some of these features are actually syntactic features, so **morpho-syntactic features** should be a better term…
- We can conceive this annotation field as a subset of the UPOS
  - For instance verbs (VERB) can be better described with the verbal form (VerbForm=) feature as Finite Verbs (Fin), Participles (Part), Gerund(ive)s (Ger), …
- This is again something working at the **language-specific level** but with **a universal set of features.**

UNIVERSAL DEPENDENCIES: UNIVERSAL MORPHOLOGICAL FEATURES (FEATS)

Universal Morphological features

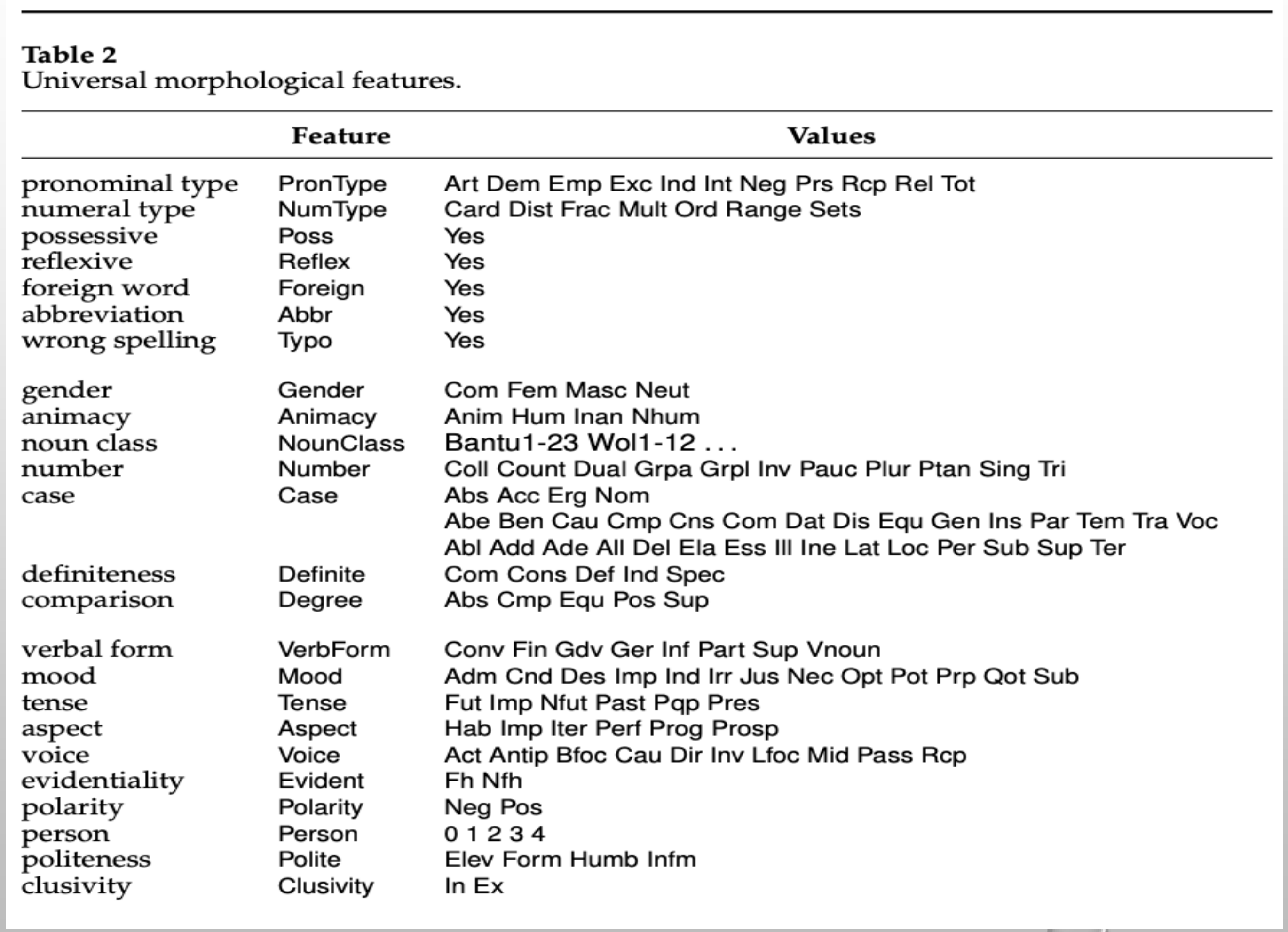


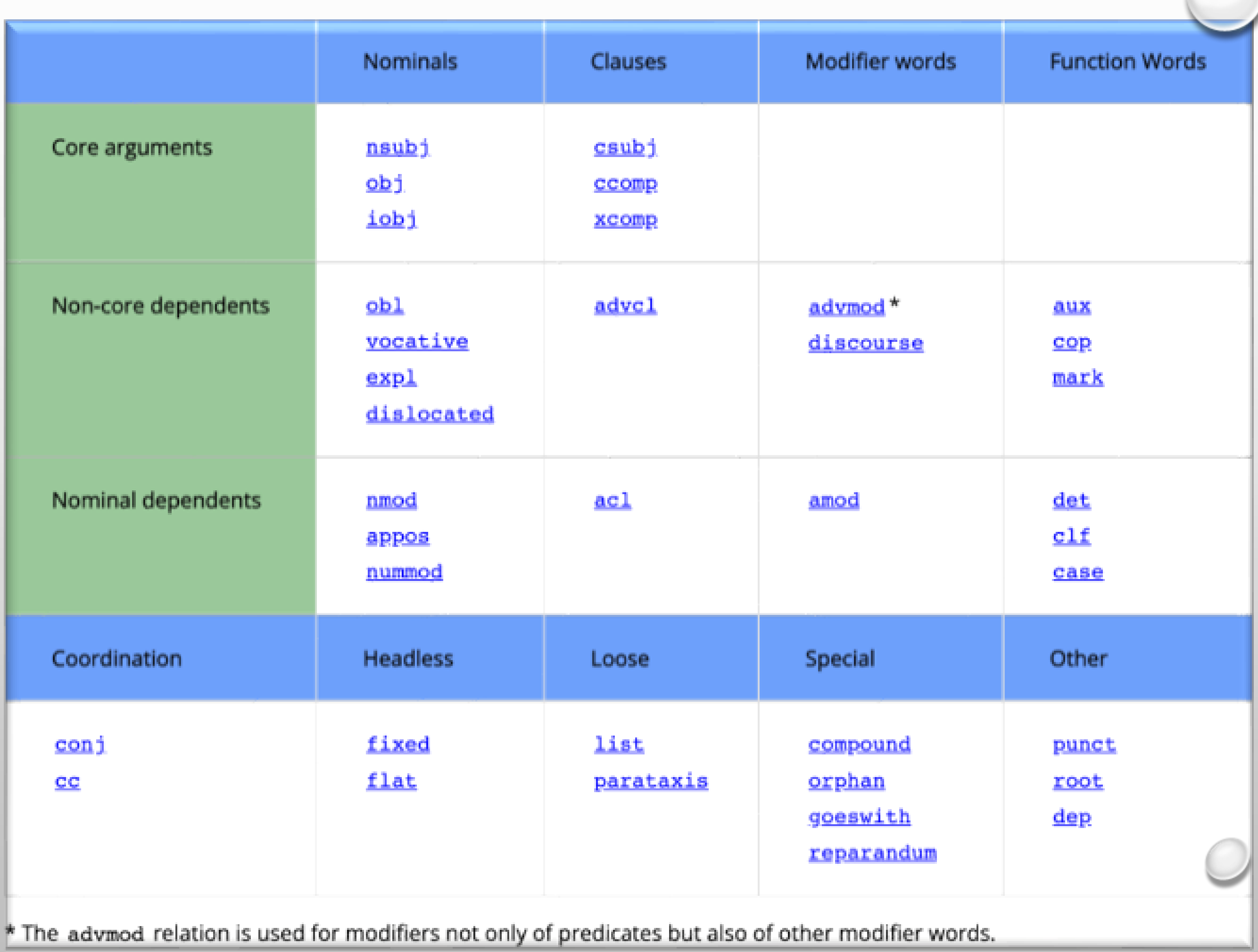
Table 2
Universal morphological features.

	Feature	Values
pronominal type	PronType	Art Dem Emp Exc Ind Int Neg Prs Rcp Rel Tot
numeral type	NumType	Card Dist Frac Mult Ord Range Sets
possessive	Poss	Yes
reflexive	Reflex	Yes
foreign word	Foreign	Yes
abbreviation	Abbr	Yes
wrong spelling	Typo	Yes
gender	Gender	Com Fem Masc Neut
animacy	Animacy	Anim Hum Inan Nhum
noun class	NounClass	Bantu1-23 Wol1-12 . . .
number	Number	Coll Count Dual Grpa Grpl Inv Pauc Plur Ptan Sing Tri
case	Case	Abs Acc Erg Nom Abe Ben Cau Cmp Cns Com Dat Dis Equ Gen Ins Par Tem Tra Voc Abl Add Ade All Del Ela Ess Ill Ine Lat Loc Per Sub Sup Ter
definiteness	Definite	Com Cons Def Ind Spec
comparison	Degree	Abs Cmp Equ Pos Sup
verbal form	VerbForm	Conv Fin Gdv Ger Inf Part Sup Vnoun
mood	Mood	Adm Cnd Des Imp Ind Irr Jus Nec Opt Pot Prp Qot Sub
tense	Tense	Fut Imp Nfut Past Pqp Pres
aspect	Aspect	Hab Imp Iter Perf Prog Prosp
voice	Voice	Act Antip Bfoc Cau Dir Inv Lfoc Mid Pass Rcp
evidentiality	Evident	Fh Nfh
polarity	Polarity	Neg Pos
person	Person	0 1 2 3 4
politeness	Polite	Elev Form Humb Infm
clusivity	Clusivity	In Ex

https://universaldependencies.org/u/feat/all.html

UNIVERSAL DEPENDENCIES: RELATIONS (HEAD+DEPREL)

- UPOS and Morphological Features ‘work’ without any other tokens, describing only some features of the annotated token;
- UD Relations, as the name implies, need exactly two tokens to work: the annotated token and its head;
- The only element without a head is the root token, which is unique to each sentence and the mother of all other tokens.
- Two fields/columns:
  - **Head**: ID of the head of the token;
  - **Deprel**: UD offers 37 Relations to describe the relation between the token and its head.



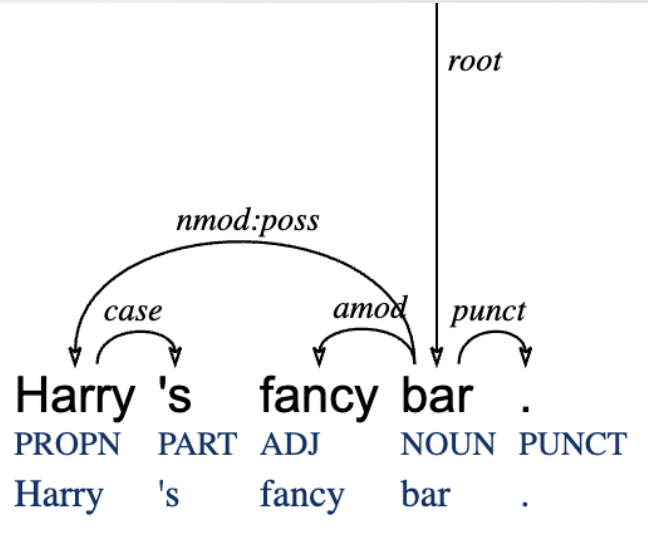
	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	Headless	Loose	Special	Other
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

* The advmod relation is used for modifiers not only of predicates but also of other modifier words.

UNIVERSAL DEPENDENCIES: RELATIONS SOME EXAMPLES

Nominal modification (nmod: a relation between two nouns), adpositions (case: syntactic case marking) and adjectival modification (amod: modification by adjectives).

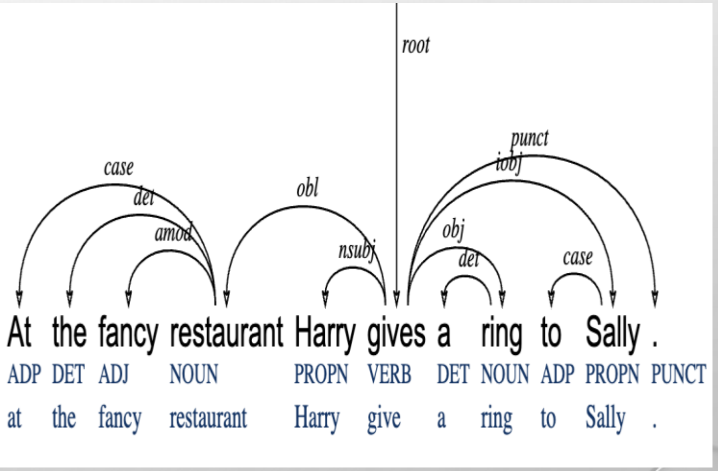
```
# text = Harry's fancy bar.
1-2     Harry's         _       _       _       _       _       _       _       _
1       Harry Harry PROPN       SG      Number=Sing     4       nmod:poss       _       _
2       's      's      PART  GEN       _       1       case    _       _
3       fancy fancy ADJ    POS  Degree=Pos 4    amod _          _
4       bar     bar     NOUN            SG-NOM  Number=Sing     0       root    _       SpaceAfter=No
5       .       .       PUNCT           Period_         4       punct _         SpaceAfter=No
```



UNIVERSAL DEPENDENCIES: RELATIONS SOME EXAMPLES

Syntactic roles: subject (nsubj), object (obj), indirect object (iobj), oblique (obl)

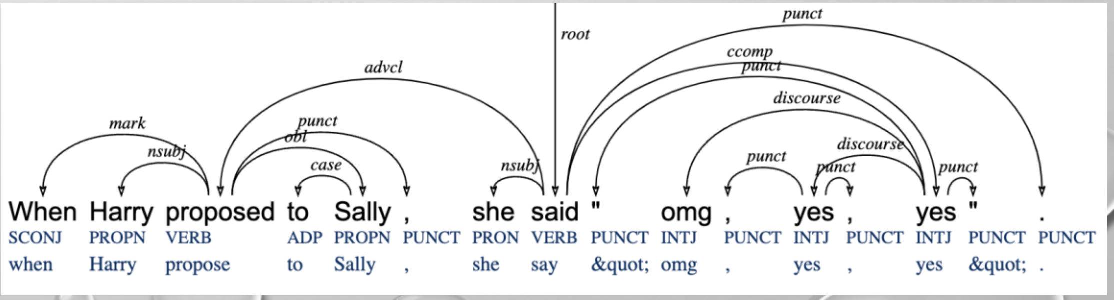
```
# sent_id = 1# text = At the fancy restaurant Harry gives a ring to Sally.
1     At     at     ADP   _     _      4     case  _     _
2     the    the    DET   DEF   Definite=Def|PronType=Art4     det   _     _
3     fancy  fancy  ADJ   POS   Degree=Pos  4     amod  _     _
4     restaurant    restaurant  NOUNSG-NOM  Number=Sing      6     obl   _     _
5     Harry  Harry  PROPN       SG-NOM      Number=Sing      6     nsubj _     _
6     gives  give   VERB  PRES  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin      0     root  _
7     a      a      DET   IND-SG      Definite=Ind|PronType=Art 8     det   _     _
8     ring   ring   NOUNSG-NOM        Number=Sing      6     obj   _     _
9     to     to     ADP   _     _      10    case  _     _
10    Sally  Sally  PROPN       SG-NOM      Number=Sing      6     iobj  _     SpaceAfter=No
11    .      .      PUNCT       Period_     6     punct _     SpaceAfter=No
# text = Harry's bar.
1     Harry  Harry  PROPN       SP    _     0     root  _     SpaceAfter=No
2     's     's     PART  PART  _     1     case  _     _
3     bar    bar    NOUNS       Gender=Masc       1     nmod  _     SpaceAfter=No
4     .      .      PUNCT       FS    _     1     punct _     SpaceAfter=No
```


root
case
det
amod
obl
nsubj
punct
iobj
obj
det
case
At the fancy restaurant Harry gives a ring to Sally .
ADP DET ADJ NOUN PROPN VERB DET NOUN ADP PROPN PUNCT
at the fancy restaurant Harry give a ring to Sally .

UNIVERSAL DEPENDENCIES: RELATIONS SOME EXAMPLES

Subordinate clauses: adverbial clauses (advcl), object clauses (ccomp)

```
# text = When Harry proposed to Sally, she said "omg, yes, yes".
1     When    when    SCONJ_      _       3       mark    _       _
2     Harry   Harry   PROPN       SG-NOM  Number=Sing 3   nsubj   _       _
3     proposed        propose     VERB    PAST    Mood=Ind|Tense=Past|VerbForm=Fin        8       advcl   _       _
4     to      to      ADP     _       _       5       case    _       _
5     Sally   Sally   PROPN SG-NOM    Number=Sing 3   obl     _       SpaceAfter=No
6     ,       ,       PUNCT Comma_    3       punct   _       _
7     she     she     PRON  PERS-SG-NOM       Case=Nom|Gender=Fem|Number=Sing|Person=3|PronType=Prs 8       nsubj   _
      _
8     said    say     VERB  PAST    Mood=Ind|Tense=Past|VerbForm=Fin        0       root    _       _
9     "       &quot; PUNCT Quote _   14      punct   _       SpaceAfter=No
10    omg     omg     INTJ  _       _       14      discourse       _       SpaceAfter=No
11    ,       ,       PUNCT Comma_  12      punct   _       _
12    yes     yes     INTJ  _       Polarity=Pos    14      discourse       _       SpaceAfter=No
13    ,       ,       PUNCT Comma_  12      punct   _       _
14    yes     yes     INTJ  _       Polarity=Pos    8       ccomp   _       SpaceAfter=No
15    "       &quot; PUNCT Quote _   14      punct   _       SpaceAfter=No
16    .       .       PUNCT Period _  8       punct   _       SpaceAfter=No
```



punct
root
ccomp
punct
advcl
discourse
mark
punct
obl
nsubj
case
nsubj
punct
discourse
punct
punct
When Harry proposed to Sally , she said " omg , yes , yes " .
SCONJ PROPN VERB ADP PROPN PUNCT PRON VERB PUNCT INTJ PUNCT INTJ PUNCT INTJ PUNCT PUNCT
when Harry propose to Sally , she say &quot; omg , yes , yes " .

UNIVERSAL DEPENDENCIES: CONLL-U FILES

From https://universaldependencies.org/format.html:

”Annotations are encoded in plain text files (UTF-8, normalized to NFC, using only the LF character as line break, including an LF character at the end of file) with three types of lines:

Word lines containing the annotation of a word/token/node in 10 fields separated by single tab characters; see below.

Blank lines marking sentence boundaries. The last line of each sentence is a blank line.

Sentence-level comments starting with hash (#). Comment lines occur at the beginning of sentences, before word lines.”

EXPERIMENTING WITH MINICIEP+

HOW TO WORK WITH MINICIEP+

- In its most basic form, miniciep+ is a collection of text (txt) files using the UTF-8 encoding, so you can explore it using a simple text editor. However, it might be not so useful, as (i) you miss any form of annotation and (ii) you cannot perform elaborate queries.
- Enter the CoNLL-U files, which are the UD-parsed version
- With or without its annotations, miniciep+ can be encoded in Corpus Query Processors such as CWB (Corpus WorkBench) or Sketch Engine, allowing you to perform complex queries.
- In this workshop, we will focus on an alternative way of exploring corpora, learning the art of extracting data from CoNLL-U files using Python scripts and storing results in comma-separated value (CSV) files for further analyses.