

UD Guidelines and Validation: How It Works and Evolves

Daniel Zeman

📅 February 7, 2024



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

Outline

- **NOT** intro to what UD is (→ Paris 2023)
- **NOT** a step-by-step demo of new treebank (→ webinar 2023)
- **BUT** something in between

- **NOT** intro to what UD is (→ Paris 2023)
- **NOT** a step-by-step demo of new treebank (→ webinar 2023)

- **BUT** something in between

- History of UD guidelines, future?
- Modular requirements
- Validation infrastructure

- Key points in guidelines
 - UPOS vs. DEPREL
 - Nominals vs. clauses vs. other
 - Core vs. oblique
 - Auxiliaries
 - Wordhood

Guidelines Approved by the Core Group



Guidelines: History and Stability

- **Stability** \Rightarrow key to success
 - Conservative approach to changing *universal* guidelines
- 2014: UD **v1** guidelines
 - Releases 1.0 (January 2015) to 1.4 (November 2016)
- end of 2016: UD **v2** guidelines
 - Releases 2.0 (March 2017) and onwards (2.13 in November 2023)

Guidelines: History and Stability

- **Stability** ⇒ key to success
 - Conservative approach to changing *universal* guidelines
- 2014: UD **v1** guidelines
 - Releases 1.0 (January 2015) to 1.4 (November 2016)
- end of 2016: UD **v2** guidelines
 - Releases 2.0 (March 2017) and onwards (2.13 in November 2023)

- Documentation can and should change/improve
- But changes are **clarifications**, not negation of a previous policy

Guidelines: History and Stability

- **Stability** ⇒ key to success
 - Conservative approach to changing *universal* guidelines
- 2014: UD **v1** guidelines
 - Releases 1.0 (January 2015) to 1.4 (November 2016)
- end of 2016: UD **v2** guidelines
 - Releases 2.0 (March 2017) and onwards (2.13 in November 2023)

- Documentation can and should change/improve
- But changes are **clarifications**, not negation of a previous policy
- 2022: **amendments** may negate a previous policy (incremental changes rather than UD v3)
 - Approved by UD Core Group (meets once a month)
 - Inspired by **GitHub issues**

Language Specific Guidelines

- Language-specific guidelines may change between releases

Language Specific Guidelines

- Language-specific guidelines may change between releases
- They must meet all requirements of the current universal guidelines
- They must be well documented

Language Specific Guidelines


- Language-specific guidelines may change between releases
- They must meet all requirements of the current universal guidelines
- They must be well documented
- **One set of guidelines per language** (not per treebank!)
 - Possible tension between treebank providers
 - Treebank-specific deviations are not supposed to occur
 - But if they do, they should be documented
 - Still better than a “secret” deviation

Validation

- Official **validation script** (Python) – tools repo on GitHub
 - Originally by Filip Ginter and Sampo Pyysalo
 - Only low-level tests (10 columns, empty lines...)
 - Online reports on UD web
 - In addition “syntax-oriented tests” (not mandatory; configured and displayed on the web)




Validation

- Official **validation script** (Python) – tools repo on GitHub
 - Originally by Filip Ginter and Sampo Pyysalo
 - Only low-level tests (10 columns, empty lines...)
 - Online reports on UD web
 - In addition “syntax-oriented tests” (not mandatory; configured and displayed on the web)
- Alternative: **Udapi** (by Martin Popel) for the syntax tests 
 - <https://udapi.github.io/>
 - Bugs were reflected in treebank star rating but did not prevent release



Validation

- Official **validation script** (Python) – tools repo on GitHub
 - Originally by Filip Ginter and Sampo Pyysalo
 - Only low-level tests (10 columns, empty lines...)
 - Online reports on UD web
 - In addition “syntax-oriented tests” (not mandatory; configured and displayed on the web)
- Alternative: **Udapi** (by Martin Popel) for the syntax tests 
 - <https://udapi.github.io/>
 - Bugs were reflected in treebank star rating but did not prevent release
- Since 2019 validator maintained by me
- Morphology and syntax tested in the validator
- New test makes data invalid → grace period
- Online report + **check_files.pl**



- 1 Technical level, backbone file format
 - E.g. Line must have 10 TAB-separated non-empty fields

Levels of Validity

- ① Technical level, backbone file format
 - E.g. Line must have 10 TAB-separated non-empty fields
- ② Low-level UD-specific tests
 - E.g. The UPOS column contains one of 17 known UPOS tags

Levels of Validity

- 1 Technical level, backbone file format
 - E.g. Line must have 10 TAB-separated non-empty fields
- 2 Low-level UD-specific tests
 - E.g. The UPOS column contains one of 17 known UPOS tags
- 3 Universal guidelines, syntax oriented
 - E.g. conj relation goes left to right

Levels of Validity

- 1 Technical level, backbone file format
 - E.g. Line must have 10 TAB-separated non-empty fields
- 2 Low-level UD-specific tests
 - E.g. The UPOS column contains one of 17 known UPOS tags
- 3 Universal guidelines, syntax oriented
 - E.g. conj relation goes left to right
- 4 Language-specific formal tests
 - E.g. Known Feature=Value pairs in the language

Levels of Validity

- 1 Technical level, backbone file format
 - E.g. Line must have 10 TAB-separated non-empty fields
- 2 Low-level UD-specific tests
 - E.g. The UPOS column contains one of 17 known UPOS tags
- 3 Universal guidelines, syntax oriented
 - E.g. conj relation goes left to right
- 4 Language-specific formal tests
 - E.g. Known Feature=Value pairs in the language
- 5 Language-specific guidelines
 - E.g. Lemma of **AUX** or cop must be registered

Language Specific Lists and Documentation

- One-page language documentation mandatory!
 - As many pages as needed possible

Language Specific Lists and Documentation

- One-page language documentation mandatory!
 - As many pages as needed possible
- Language-specific documentation of features
 - Mandatory for each extra feature
 - Prescribed format of the doc page!
 - Mandatory for universal feature if there are extra values
 - Such page overrides the global one \Rightarrow undocumented universal values are no longer allowed!
 - Optional for other universal features

Language Specific Lists and Documentation

- **One-page language documentation** mandatory!
 - As many pages as needed possible
- Language-specific **documentation of features**
 - Mandatory for each extra feature
 - Prescribed format of the doc page!
 - Mandatory for universal feature if there are extra values
 - Such page overrides the global one \Rightarrow undocumented universal values are no longer allowed!
 - Optional for other universal features
- Register permitted UPOS-feature-value combinations
 - `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl`

Language Specific Lists and Documentation

- **One-page language documentation** mandatory!
 - As many pages as needed possible
- Language-specific **documentation of features**
 - Mandatory for each extra feature
 - Prescribed format of the doc page!
 - Mandatory for universal feature if there are extra values
 - Such page overrides the global one \Rightarrow undocumented universal values are no longer allowed!
 - Optional for other universal features
- Register permitted UPOS-feature-value combinations
 - `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl`
- Language-specific **relation subtypes**

Language Specific Lists and Documentation

- **One-page language documentation** mandatory!
 - As many pages as needed possible
- Language-specific **documentation of features**
 - Mandatory for each extra feature
 - Prescribed format of the doc page!
 - Mandatory for universal feature if there are extra values
 - Such page overrides the global one \Rightarrow undocumented universal values are no longer allowed!
 - Optional for other universal features
- Register permitted UPOS-feature-value combinations
 - https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl
- Language-specific **relation subtypes**
- Lists of **auxiliaries** and copulas (documented)

Language Specific Lists and Documentation

- **One-page language documentation** mandatory!
 - As many pages as needed possible
- Language-specific **documentation of features**
 - Mandatory for each extra feature
 - Prescribed format of the doc page!
 - Mandatory for universal feature if there are extra values
 - Such page overrides the global one \Rightarrow undocumented universal values are no longer allowed!
 - Optional for other universal features
- Register permitted UPOS-feature-value combinations
 - `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl`
- Language-specific **relation subtypes**
- Lists of **auxiliaries** and copulas (documented)
- Enhanced dependencies: Case-enhanced relations

Language Specific Lists and Documentation

- **One-page language documentation** mandatory!
 - As many pages as needed possible
- Language-specific **documentation of features**
 - Mandatory for each extra feature
 - Prescribed format of the doc page!
 - Mandatory for universal feature if there are extra values
 - Such page overrides the global one \Rightarrow undocumented universal values are no longer allowed!
 - Optional for other universal features
- Register permitted UPOS-feature-value combinations
 - `https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/langspec/specify_feature.pl`
- Language-specific **relation subtypes**
- Lists of **auxiliaries** and copulas (documented)
- Enhanced dependencies: Case-enhanced relations
- Permitted words with spaces (regular expressions)

Features

Lexical	Inflectional (“Nominal”)	Inflectional (“Verbal, Pronominal”)
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflect	Number	Aspect
Foreign	Case	Voice
	Definite	Evident
	Deixis	Polarity
Abbr	DeixisRef	Person
Typo	Degree	Polite
		Clusivity

- 26 features, each with a number of possible *values*
- Languages select relevant features
- May add language-specific features or values

Language Specific Lists and Documentation

- Universal features/dependencies
 - Defined and documented globally
 - No need to document them for each language, but possible (with lang-spec examples)

- Language-specific features/dependency subtypes
 - Must be documented for each language
 - Even if already documented for another language

Language Specific Lists and Documentation

- Universal features/dependencies
 - Defined and documented globally
 - No need to document them for each language, but possible (with lang-spec examples)
- **Gray zone:** not universal, but used widely
 - Some features/subtypes are documented globally...
 - ... although they officially do not count as “universal”
 - Some features silently incorporated in the universal pool
- Language-specific features/dependency subtypes
 - Must be documented for each language
 - Even if already documented for another language

IMPORTANT: Do not make other people's treebanks invalid!

- Error in documentation page \Rightarrow the feature is no longer available in the language
- You don't use a feature value for, say, nouns?
- Don't uncheck it unless you are sure that nobody else uses it!
 - Even if it is clearly wrong, talk to the maintainers of the other treebank first, or fix the data if the treebank is no longer maintained

Future Perspectives

- New lists / language-specific tests in the validator

Future Perspectives

- New lists / language-specific tests in the validator
- Possible tightening of MISC and # comments, documenting their options

Future Perspectives

- New lists / language-specific tests in the validator
- Possible tightening of MISC and # comments, documenting their options
- **Backward compatibility** is important in low-level processing
 - \Rightarrow do not touch CoNLL-U :-)
 - The 17 UPOS tags are also carved in stone

Future Perspectives

- New lists / language-specific tests in the validator
- Possible tightening of MISC and # comments, documenting their options
- **Backward compatibility** is important in low-level processing
 - \Rightarrow do not touch CoNLL-U :-)
 - The 17 UPOS tags are also carved in stone
- Enhanced UD guidelines are less developed and less widely used \Rightarrow I can imagine changes there

Future Perspectives

- New lists / language-specific tests in the validator
- Possible tightening of MISC and # comments, documenting their options
- **Backward compatibility** is important in low-level processing
 - \Rightarrow do not touch CoNLL-U :-)
 - The 17 UPOS tags are also carved in stone
- Enhanced UD guidelines are less developed and less widely used \Rightarrow I can imagine changes there
- Eventually UD v3 might actually happen
 - There is a wish list of changes that cannot be amendments
 - Renaming a main relation type
 - Possibly allowing CoNLL-U-Plus, with some restrictions

CoNLL-U Format

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Es	es	PRON	-	-	2	nsubj	-	-
2	unterscheidet	unterscheiden	VERB	-	-	0	root	-	-
3	sich	sich	PRON	-	-	2	expl:pv	-	-
4-5	vom	-	-	-	-	-	-	-	-
4	von	von	ADP	-	-	7	case	-	-
5	dem	der	DET	-	-	7	det	-	-
6	westlichen	westlich	ADJ	-	-	7	amod	-	SpaceAfter=No
7	Teil	Teil	NOUN	-	-	2	obl	-	-
8	des	der	DET	-	-	9	det	-	-
9	Landes	Land	NOUN	-	-	7	nmod	-	SpaceAfter=No
10	.	.	PUNCT	-	-	2	punct	-	-

- **UPOS**, **HEAD** and **DEPREL** must be manual
- **LEMMA**, **FEATS**, **DEPS** can be omitted or automatic (but still must validate)
- **XPOS** can be almost anything – UD does not care
- **MISC**: UD slightly cares but the columns takes everything that does not fit elsewhere

- UPOS is largely syntactic, yet not the same as DEPREL
 - Complete predictability UPOS \Rightarrow DEPREL would be uninteresting
 - UPOS: Typical behavior
 - DEPREL: Actual function at this position

- UPOS is largely syntactic, yet not the same as DEPREL
 - Complete predictability UPOS \Rightarrow DEPREL would be uninteresting
 - UPOS: Typical behavior
 - DEPREL: Actual function at this position
- Nevertheless, there is correlation
- Some UPOS-DEPREL combinations are invalid!

Dependents of Clauses (Verbal or Not)

	Nominal	Clausal	Modifier	Function
Core	nsubj obj iobj	csubj ccomp xcomp		
Non-Core	obl vocative dislocated expl	advcl	advmod discourse	aux cop mark

Dependents of Adjectives and Adverbs (Predicative or Not)

	Nominal	Clausal	Modifier
Core	(obj) (iobj)	ccomp xcomp	
Non-Core	obl expl	advcl	advmod

Dependents of Nominals

Nominal

nmod
(appos)
(compound)
(flat)

Clausal

acl

Modifier

amod
nummod

Function

det
case
clf

Nominalized Verbs

AS DEPENDENT:

- Did you tag it **VERB VerbForm=Vnoun**?
 - ⇒ **clause!** (csubj, ccomp, advcl, acl)

AS HEAD:

- Did you tag it **VERB VerbForm=Vnoun**?
 - ⇒ dependents of clauses (nsubj, csubj, obj, iobj, ccomp, xcomp, obl, advcl...)

Nominalized Verbs

AS DEPENDENT:

- Did you tag it **VERB** **VerbForm=Vnoun**?
 - ⇒ **clause!** (csubj, ccomp, advcl, acl)
- Did you tag it **NOUN** **VerbForm=Vnoun**?
 - Is it a predicate?
 - ⇒ **clause!** (see above)

AS HEAD:

- Did you tag it **VERB** **VerbForm=Vnoun**?
 - ⇒ dependents of clauses (nsubj, csubj, obj, iobj, ccomp, xcomp, obl, advcl...)
- Did you tag it **NOUN** **VerbForm=Vnoun**?
 - Is it a predicate?
 - ⇒ dependents of **both clauses and nominals** (see above and below)

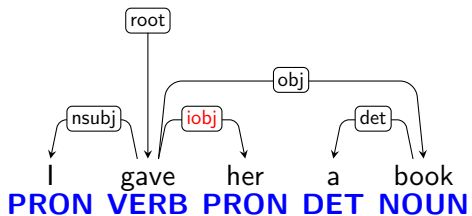
AS DEPENDENT:

- Did you tag it **VERB** **VerbForm=Vnoun**?
 - ⇒ **clause!** (csubj, ccomp, advcl, acl)
- Did you tag it **NOUN** **VerbForm=Vnoun**?
 - Is it a predicate?
 - ⇒ **clause!** (see above)
 - No?
 - ⇒ **nominal!** (nsubj, obj, obl, nmod...)

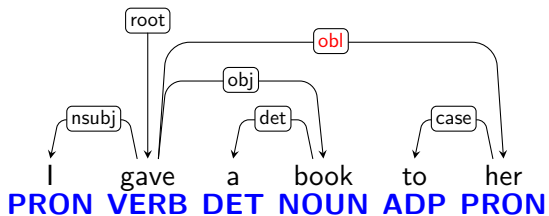
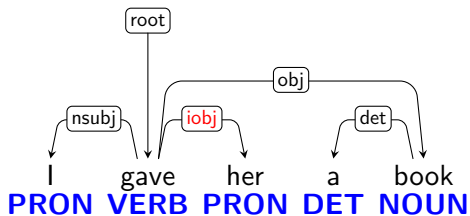
AS HEAD:

- Did you tag it **VERB** **VerbForm=Vnoun**?
 - ⇒ dependents of clauses (nsubj, csubj, obj, iobj, ccomp, xcomp, obl, advcl...)
- Did you tag it **NOUN** **VerbForm=Vnoun**?
 - Is it a predicate?
 - ⇒ dependents of **both clauses and nominals** (see above and below)
 - No?
 - ⇒ dependents of nominals (nmod, amod, acl...)

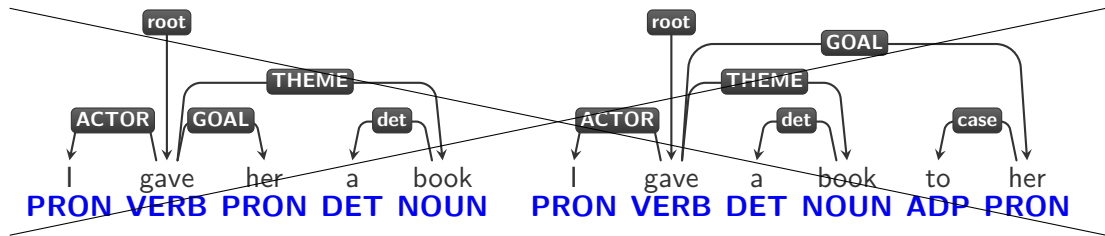
Core vs. Oblique: Information Packaging



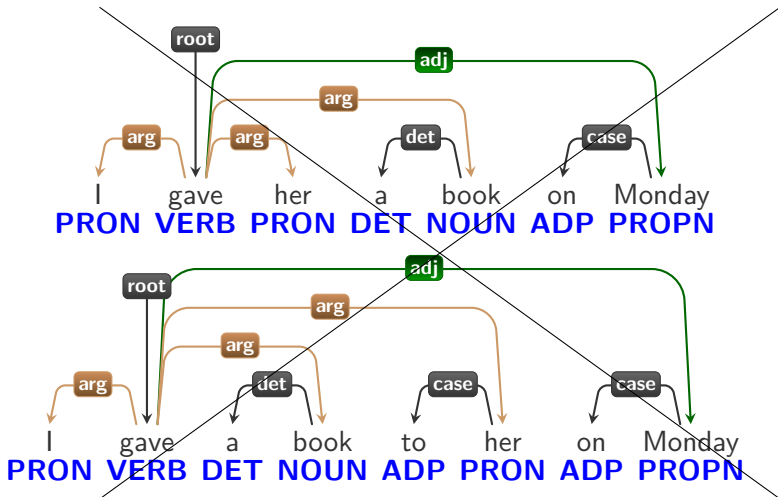
Core vs. Oblique: Information Packaging



UD is NOT about Semantic Roles!



UD Avoids Argument-Adjunct Distinction!



Core vs. Oblique

- English (simplified):
 - Bare noun phrase \Rightarrow core argument (nsubj, obj, iobj)
 - Prepositional phrase \Rightarrow oblique argument or adjunct (obl)

Core vs. Oblique

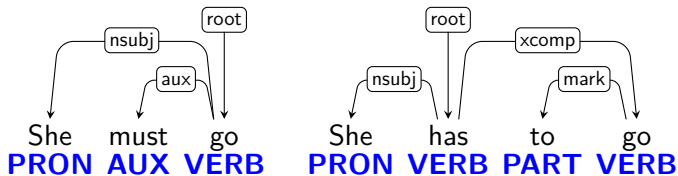
- English (simplified):
 - Bare noun phrase \Rightarrow core argument (nsubj, obj, iobj)
 - Prepositional phrase \Rightarrow oblique argument or adjunct (obl)
- Other languages: not necessarily!
 - Japanese and Tagalog use adpositions with core arguments
 - Spanish and Hindi have adposition for animate direct objects

Core vs. Oblique

- English (simplified):
 - Bare noun phrase \Rightarrow core argument (nsubj, obj, iobj)
 - Prepositional phrase \Rightarrow oblique argument or adjunct (obl)
- Other languages: not necessarily!
 - Japanese and Tagalog use adpositions with core arguments
 - Spanish and Hindi have adposition for animate direct objects
- Clash with traditional terminology
 - Grammars of German, Czech etc. define **prepositional objects**
 - But these are not necessarily core...
 - **Morphological cases**: dative is oblique in most Indo-European languages (but may be core elsewhere)
 - Same for genitive, locative, ablative, instrumental

Auxiliaries (including Copulas)

- Just because traditional grammar calls it auxiliary does not mean it is a UD auxiliary!
- Just because it has a modal meaning does not mean it is a UD auxiliary!
 - It should have grammatical function comparable to morphological features of **Tense, Aspect, Mood, Voice, Evident**... Or it may bear verbal agreement with core arguments.
 - Ideally, there should be language-specific tests that distinguish it from other verbs that take verbs as complements.



- **Alternatives to aux:** xcomp, compound

Word Segmentation

Let's go to the sea.

Vámonos al mar . Vamos nos a el mar .
VERB? X NOUN PUNCT VERB PRON ADP DET NOUN PUNCT

- **Syntactic word** vs. orthographic word
- **Multi-word tokens**
- Two-level scheme:
 - Tokenization (low level, punctuation, concatenative)
 - Word segmentation (higher level, not necessarily concatenative)

Recoverability: CoNLL-U Format

text = Vámonos al mar.

text_en = Let's go to the sea.

ID	FORM	LEMMA	UPOS	...	HEAD	_ MISC
1-2	Vámonos	—	—	...	—	—
1	Vamos	ir	VERB	...	0	root
2	nos	nosotros	PRON	...	1	obj
3-4	al	—	—	...	—	—
3	a	a	ADP	...	5	case
4	el	el	DET	...	5	det
5	mar	mar	NOUN	...	1	obl
6	.	.	PUNCT	...	1	punct

SpaceAfter=No

Recoverability: CoNLL-U Format

text = Vámonos al mar.

text_en = Let's go to the sea.

ID	FORM	LEMMA	UPOS	...	HEAD	_	MISC
1-2	Vámonos	—	—	...	—	—	— —
1	Vamos	ir	VERB	...	0	root	— —
2	nos	nosotros	PRON	...	1	obj	— —
3-4	al	—	—	...	—	—	— —
3	a	a	ADP	...	5	case	— —
4	el	el	DET	...	5	det	— —
5-6	mar.	—	—	...	—	—	— —
5	mar	mar	NOUN	...	1	obl	— —
6	.	.	PUNCT	...	1	punct	— —

Vietnamese: Words with Spaces

All the concrete country roads are the result of...

Tất cả	đường	bê tông	nội đồng	là	thành quả	...
All	road	concrete	country	is	achievement	...
PRON	NOUN	NOUN	NOUN	AUX	NOUN	PUNCT

- Spaces delimit monosyllabic morphemes, not words.
- Multiple syllables without space occur in loanwords (*bê tông*).
- Spaces are allowed to occur word-internally in Vietnamese UD.

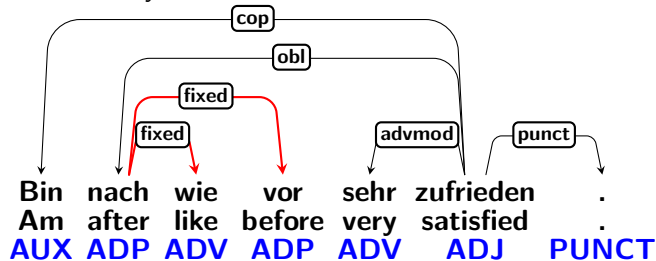
Numbers with Spaces

#	text = Il touche environ 100 000 sesterces par an.						
1	Il	il	PRON	...	2	nsubj	--
2	touche	toucher	VERB	...	0	root	--
3	environ	environ	ADV	...	4	advmod	--
4	100 000	100 000	NUM	...	5	nummod	--
5	sesterces	sesterce	NOUN	...	2	obj	--
6	par	par	ADP	...	7	case	--
7	an	an	NOUN	...	2	obl	_ SpaceAfter=No
8	.	.	PUNCT	...	2	punct	--

Fixed Expressions

One syntactic word spans several orthographic words?

I am still very satisfied.



Part-of-Speech Tags

<http://universaldependencies.org/u/pos/index.html>

Open		Closed		Other	
NOUN	common noun	PRON	pronoun	PUNCT	punctuation
PROPN	proper noun	DET	determiner	SYM	symbol
VERB	verb	AUX	auxiliary	X	unknown
ADJ	adjective	NUM	numeral		
ADV	adverb	ADP	adposition		
INTJ	interjection	SCONJ	subordinator		
		CCONJ	coordinator		
		PART	particle		

- Taxonomy of 17 universal POS tags
- All languages use the same inventory
 - Not all tags have to be used by all languages
 - Need extensions? Use features!

Multiword Expressions

Relation	Examples
fixed	<i>as well, by and large, according to, more than</i>
flat	<i>president Havel, New York, four thousand</i>
compound	<i>phone book, dress up</i>
goeswith	<i>notwith standing, with out</i>

- UD annotation **almost** does not permit “words with spaces”
 - Multiword expressions are analyzed using special relations
 - The **fixed**, **flat** and **goeswith** relations are always head-initial
 - The **compound** relation reflects the internal structure
- Words with spaces allowed in exceptional cases:
 - Vietnamese (spaces delimit syllables, not words)
 - Numbers (“1 000 000”)
 - Possibly other approved cases, e.g. multi-word abbreviations

Language-specific Relation Subtypes

- Language-specific relations are **subtypes** of universal relations added to capture important phenomena
- Subtyping permits us to “back off” to universal relations

Language-specific Relation Subtypes

Relation	Explanation
acl:relcl	Relative clause (the boy who lived)
compound:prt	Verb particle (dress up)
nmod:poss	Possessive nominal (Mary 's book)
obl:agent	Agent in passive (saved by the bell)
cc:preconj	Preconjunction (both ... and)
det:predet	Predeterminer (all those ...)