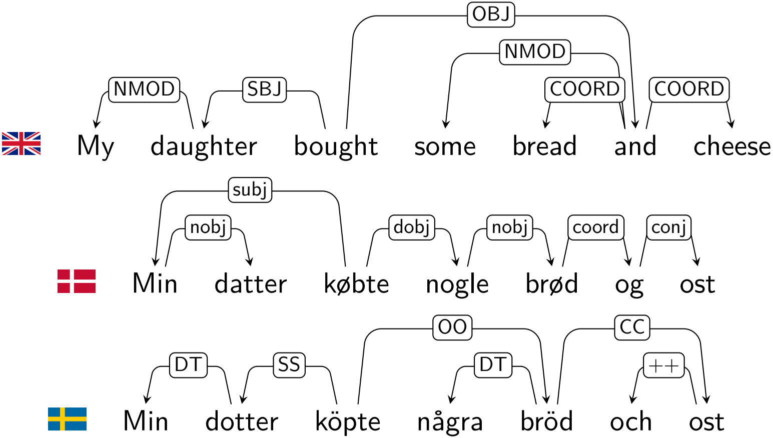
Universal Dependencies: Common Morphology and Syntax for Multiple Languages

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Jan Hajič (with a lot of Dan Zeman's slides)

Institute of Formal and Applied Linguistics & LINDAT/CLARIN
Charles University, Prague, Czech Republic
{hajic,zeman}@ufal.mff.cuni.cz
http://universaldependencies.org/



My daughter bought some bread and cheese

Min datter købte nogle brød og ost

Min dotter köpte några bröd och ost

Outline

- A Bit of History
- Goals and Requirements
- Desing Principles (and the Manning's Law)
- Morphology
- Syntax
- Word segmentation
- Some interesting phenomena - copulas, ellipsis, ...
- Current Status of Universal Dependencies
- The CoNLL 2017 Shared Task on Universal Dependencies

Universal Dependencies

http://universaldependencies.org/

Nivre Joakim et al.: Universal Dependencies v1: A Multilingual Treebank Collection. In: *Proceedings of the 10th LREC*, pp. 1659-1666, 2016

Milestones:

- 2008-05 Interset (morphological features)
- 2012-05 Google Universal POS tags
- 2012-05 HamleDT (harmonized Prague-style dependency treebanks)
- 2013-08 Google Universal Dependency Treebank
- 2014-02 Dagstuhl Seminar 14061: informal session about UD
- 2014-04 EACL Göteborg, kick-off meeting of UD, organized by J. Nivre
- 2014-05 Universal Stanford Dependencies
- 2014-10 UD guidelines version 1
- 2015-01 Released first 10 treebanks
- Every ~6 months new release
- 2016-12 UD guidelines version 2
- 2017-03 First v2 release, 70 treebanks, CoNLL Shared Task

Goals and Requirements

- Cross-linguistically consistent grammatical annotation

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de-facto standards

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de-facto standards
- Caveats:
  - Not a new linguistic theory –
    but linguistically informed and relevant

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de-facto standards
- Caveats:
  - Not a new linguistic theory –
    but linguistically informed and relevant
  - Not an ideal parsing representation –
    but useful for comparative evaluation

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de-facto standards
- Caveats:
    - Not a new linguistic theory – but linguistically informed and relevant
    - Not an ideal parsing representation – but useful for comparative evaluation
    - Not the ultimate annotation scheme – but a lightweight lingua franca

Not "Universal" in the strictly typological sense!

Design Principles

- Dependency
  - Widely used in practical NLP systems
  - Available in treebanks for many languages

Design Principles

- Dependency
  - Widely used in practical NLP systems
  - Available in treebanks for many languages
- Lexicalism
  - Basic annotation units are words – syntactic words
  - Words have morphological properties
  - Words enter into syntactic relations

Design Principles

- Dependency
  - Widely used in practical NLP systems
  - Available in treebanks for many languages
- Lexicalism
  - Basic annotation units are words – syntactic words
  - Words have morphological properties
  - Words enter into syntactic relations
- Recoverability
  - Transparent mapping from input text to word segmentation

Golden Rules

- Maximize parallelism
  - Don't annotate the same thing in different ways
  - Don't make different things look the same

Golden Rules

- Maximize parallelism
  - Don't annotate the same thing in different ways
  - Don't make different things look the same
- But don't overdo it
  - Balance: is it still the same thing?
  - Don't annotate things that are not there
  - Allow language-specific extensions

Manning's Law

The secret to understanding the design and current success of UD is to realize that the design is a very subtle compromise between approximately 6 things - UD needs to/must be:

- satisfactory on linguistic analysis grounds for individual languages.
- good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
- suitable for rapid, consistent annotation by a human annotator.
- suitable for computer parsing with high accuracy.
- easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. ... it leads us to favor traditional grammar notions and terminology.
- support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, ...).

It's easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
Some	*girls*		*nevertheless*	*praised*	*ice-cream*	*.*

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
Some	*girls*		*nevertheless*	*praised*	*ice-cream*	*.*
některý	dívka	se	nicméně	pochvalovat	zmrzlina	.

- Lemma representing the semantic content of the word

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
Some	*girls*		*nevertheless*	*praised*	*ice-cream*	*.*
některý	dívka	se	nicméně	pochvalovat	zmrzlina	.
DET	NOUN	PRON	CCONJ	VERB	NOUN	PUNCT

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
Some	*girls*		*nevertheless*	*praised*	*ice-cream*	*.*
některý	dívka	se	nicméně	pochvalovat	zmrzlina	.
DET	NOUN	PRON	CCONJ	VERB	NOUN	PUNCT
PronType=Ind Gender=Fem Number=Plur Case=Nom	Gender=Fem Number=Plur Case=Nom	PronType=Prs Reflex=Yes Case=Dat		VerbForm=Part Tense=Past Voice=Act Aspect=Imp Gender=Fem Number=Plur	Gender=Fem Number=Sing Case=Acc	

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word
- Features representing lexical and grammatical properties associated with the lemma or the particular word form

Part-of-Speech Tags

Open	***Closed***	***Other***
ADJ	*ADP*	*PUNCT*
ADV	*AUX*	*SYM*
INTJ	*CCONJ*	*X*
NOUN	*DET*	
PROPN	*NUM*	
VERB	*PART*	
	PRON	
	SCONJ	

- Taxonomy of 17 universal part-of-speech tags, based on the Google Universal Tagset (Petrov et al., 2012)
- All languages use the same inventory, but not all tags have to be used by all languages

Features (morphology++)

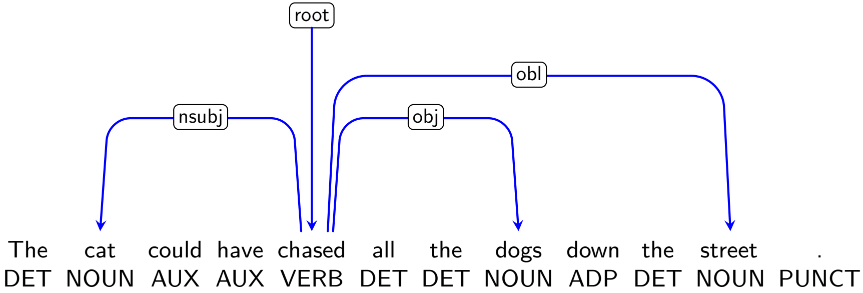
Lexical	***Inflectional (Nominal)***	***Inflectional (Verbal)***
PronType	*Gender*	*VerbForm*
NumType	*Animacy*	*Mood*
Poss	*Number*	*Tense*
Reflect	*Case*	*Aspect*
Foreign	*Definite*	*Voice*
	Degree	*Evident*
		Person
		Polite
Abbr		*Polarity*

- Standardized inventory of morphological features, based on Interset (Zeman, 2008)
- Languages select relevant features and can add language-specific features or values (with proper documentation!)

Syntax

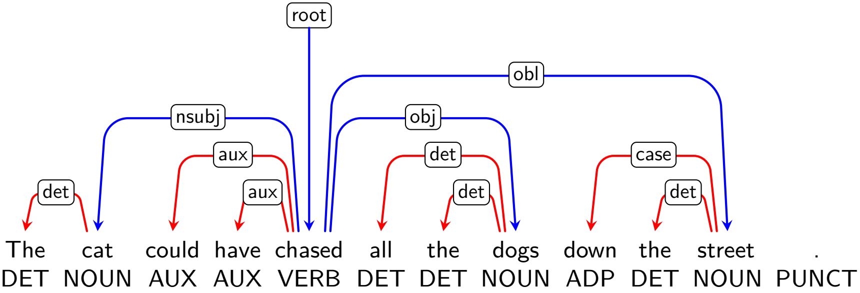
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	NOUN	AUX	AUX	VERB	DET	DET	NOUN	ADP	DET	NOUN	PUNCT

Syntax



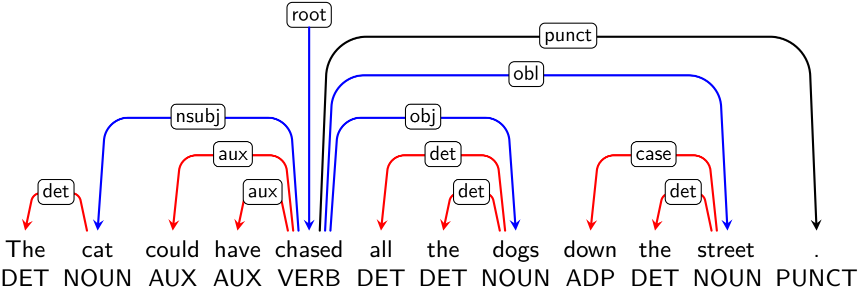
- Content words are related by dependency relations

Syntax



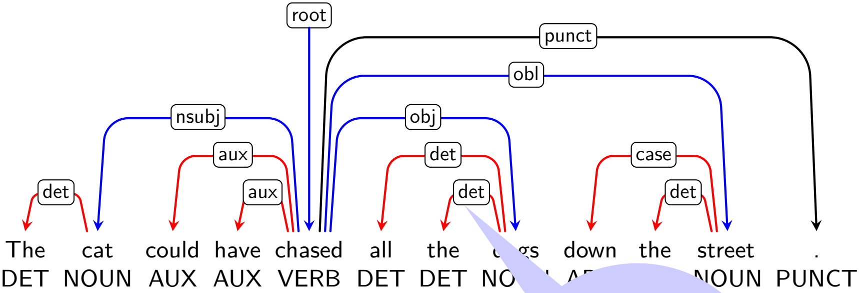
- Content words are related by dependency relations
- Function words attach to closest content words they “belong” to

Syntax

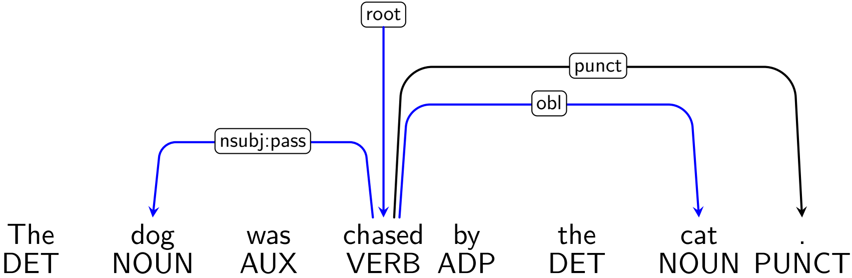

root
punct
obl
nsubj
obj
aux
det
case
aux
det
det
det
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT

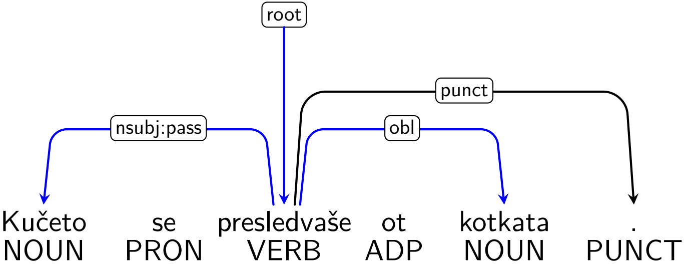
- Content words are related by dependency relations
- Function words attach to closest content words they “belong” to
- Punctuation attach to head of phrase or clause

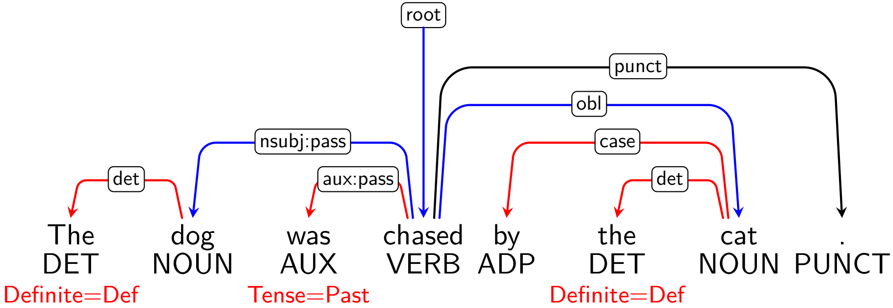
Syntax


root
punct
obl
nsubj
obj
aux
det
case
aux
det
det
det
The cat could have chased all the dogs down the street .
DET NOUN AUX AUX VERB DET DET NO A NOUN PUNCT

Not “dependency” in the strictly syntactic sense!


root
punct
obl
nsubj:pass
The DET
dog NOUN
was AUX
chased VERB
by ADP
the DET
cat NOUN
. PUNCT

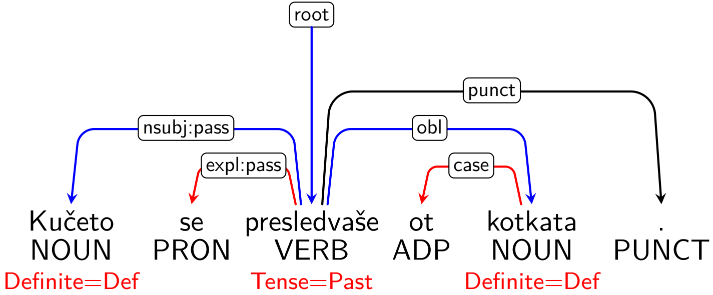

root
punct
nsubj:pass
obl
Kučeto NOUN
se PRON
presledvaše VERB
ot ADP
kotkata NOUN
. PUNCT



The dog was chased by the cat .

DET NOUN AUX VERB ADP DET NOUN PUNCT

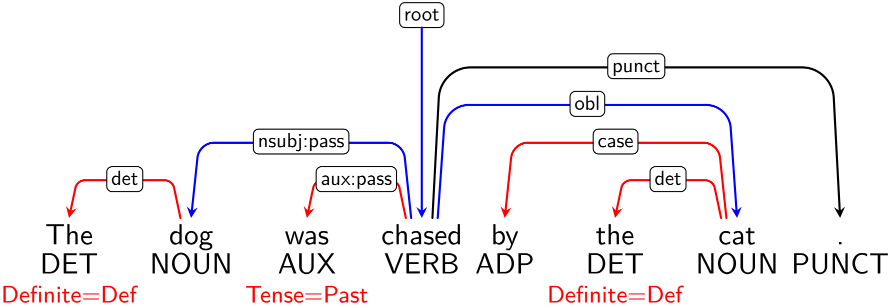
Definite=Def Tense=Past Definite=Def



Kučeto se presledvaše ot kotkata .

NOUN PRON VERB ADP NOUN PUNCT

Definite=Def Tense=Past Definite=Def



root

punct

obl

nsubj:pass

case

det

aux:pass

det

The DET Definite=Def

dog NOUN

was AUX Tense=Past

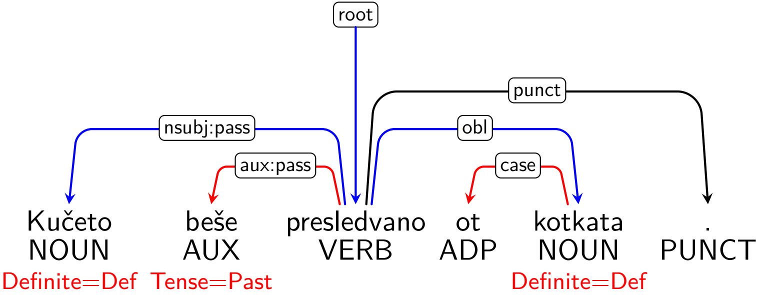
chased VERB

by ADP

the DET Definite=Def

cat NOUN

. PUNCT



root

punct

nsubj:pass

obl

aux:pass

case

Kučeto NOUN Definite=Def

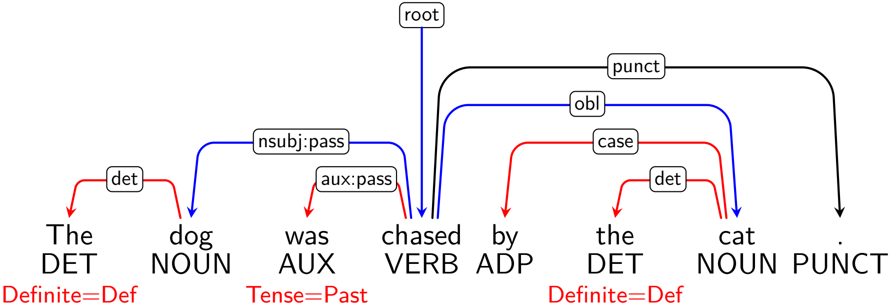
beše AUX Tense=Past

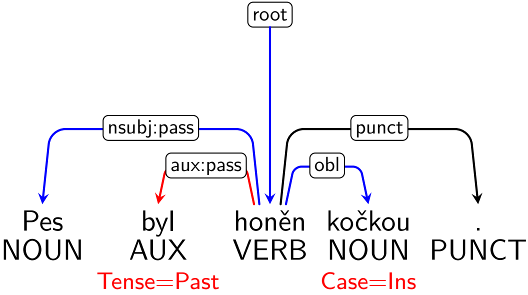
presledvano VERB

ot ADP

kotkata NOUN Definite=Def

. PUNCT


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The DET Definite=Def
dog NOUN
was AUX Tense=Past
chased VERB
by ADP
the DET Definite=Def
cat NOUN
. PUNCT


root
nsubj:pass
punct
aux:pass
obl
Pes NOUN
byl AUX Tense=Past
honěn VERB
kočkou NOUN Case=Ins
. PUNCT

Dependency Relations

- Taxonomy of 38 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
    - Language-specific subtypes may be added

Dependency Relations

- Taxonomy of 38 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
    - Language-specific subtypes may be added
- Organizing principles
    - Three types of structures: nominals, clauses, modifiers
    - Core arguments vs. other dependents (not arguments vs. adjuncts)

Core Arguments

- Easier cross-linguistically than argument-adjunct?
- Subject of intransitive verb
- Agent of transitive verb
- Patient (direct object) of transitive verb

- Indirect object? Dative only?

Core vs. Oblique Dependents

- **Core arguments:** what exactly is it?
- English:
  - *He gave John the book.* (iobj)
  - *He gave the book to John.* (obl)
- Spanish:
  - *Dio el libro a John.* (iobj)
- Czech:
  - PDT's Objs are translated mostly to obj, but there are rules to translate them to other relations if necessary (Czech Objs in PDT are more like Arguments)

Direct and Indirect Object

- Not as easy as accusative vs. dative.
- Default: obj
- Heuristics for iobj
  - *Cením si vaší pomoci.* (Gen)
    I appreciate your help.
  - *Čelíme velkým problémům.* (Dat)
    We are facing big problems.
  - *Nedisponuje takovým rozpočtem.* (Ins)
    He does not have such budget.
  - *Učí mou dceru fyziku.* (2 × Acc)
    He teaches my daughter physics.

Dependency Relations

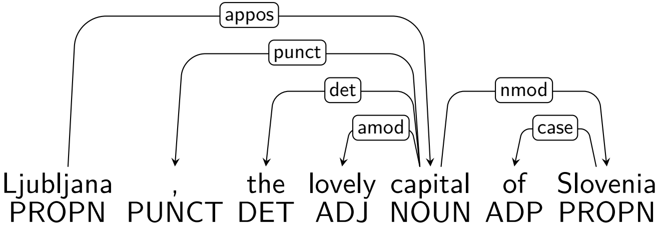
Dependents of Clausal Predicates

	Nominal	***Clausal***	***Other***
Core	*nsubj*	*csubj*	
	obj	*ccomp*	
	iobj	*xcomp*	
Non-Core	*obl*	*advcl*	*advmod*
	vocative		*aux*
	discourse		*cop*
	expl		*mark*
			punct

Dependency Relations

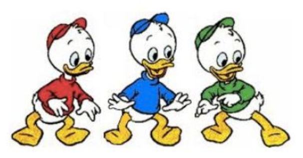
Dependents of Nominals

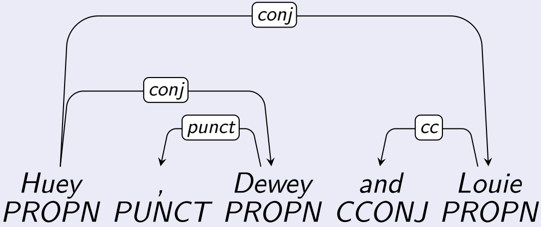
Nominal	***Clausal***	***Other***
nmod	*acl*	*amod*
appos		*det*
nummod		*case*
clf		



Dependency Relations

Coordination, modified “Stanford style”





- Coordinate structures are headed by the first conjunct
  - Subsequent conjuncts depend on it via the conj relation
  - Conjunctions depend on the next conjunct via the cc relation
  - Punctuation marks depend on the next conjunct via the punct relation

Dependency Relations

Multiword Expressions

Relation	***Examples***
fixed	*in spite of, as well as, ad hoc*
flat	*president Havel, New York, four thousand*
compound	*phone book, dress up*
goeswith	*notwith standing, with out*

- 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
  - Vietnamese (spaces delimit syllables, not words)
  - Numbers (“1 000 000”)
  - Possibly other approved cases, e.g. multi-word abbreviations

Dependency Relations

Other Relations

Relation	***Explanation***
parataxis	*Loosely linked clauses of same rank*
list	*Lists without syntactic structure*
orphan	*Orphans in ellipsis linked together*
reparandum	*Disfluency linked to (speech) repair*
foreign	*Elements within opaque stretches of code switching*
dep	*Unspecified dependency*
root	*Syntactically independent element of clause/phrase*

Language-Specific Relations

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

Language-Specific Relations

Relation	***Explanation***
acl:relcl	*Relative clause*
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 ...)*

Word Segmentation

- Must be reproducible on new data
- Surface tokens vs. syntactic words
- Chinese, Vietnamese etc.: no clues, non-trivial algorithm
- Arabic, Tamil etc.: part of morphological analysis
- Spanish, German etc.: rather limited cases of contractions
- Others: only punctuation (low-level tokenization)

Word Segmentation

Vamos	*nos*	*a*	*el*	*mar*	*.*
VERB	*PRON*	*ADP*	*DET*	*NOUN*	*PUNCT*

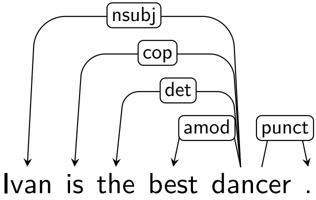
Vámonos	*al*	*mar*	*.*
VERB+PRON	*ADP+DET*	*NOUN*	*PUNCT*

Word Segmentation

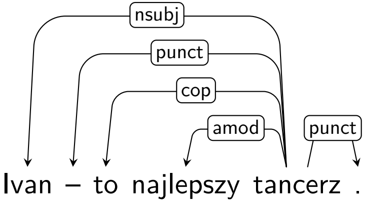
- Fusions
  - al = a + el
  - naň = na + něj
- Clitics
  - vámonos = vamos + nos
  - izmenjat'sja = izmenjat' + sja
  - potrafilibyśmy = potrafili + by + jesteśmy

Nonverbal Predicate and Copula

- Some languages use a copula verb:

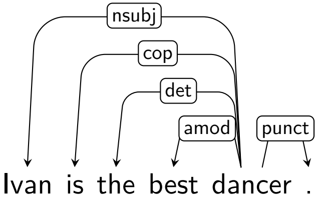


- Some languages use a copula pronoun:

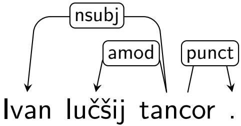


Nonverbal Predicate and Copula

- Some languages use a copula verb:

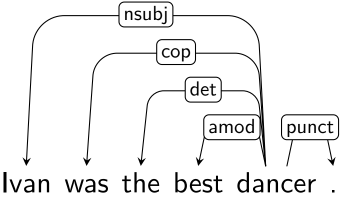


- Some languages omit the copula:

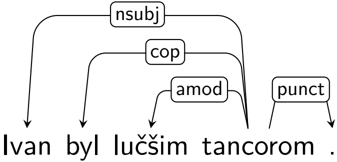


Nonverbal Predicate and Copula

- Some languages use a copula verb:

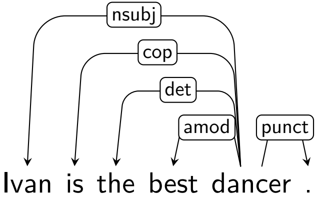


- Some languages use it only in some tenses:

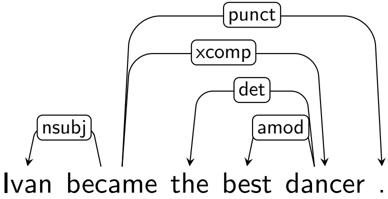


Copula Verbs: We Are Restrictive!

- *To be* is copula:

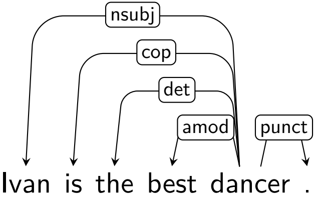


- *To become* is not copula:

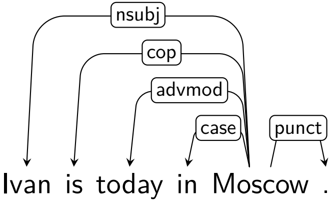


Once Copula, Always Copula!

- This is parallel with Russian:

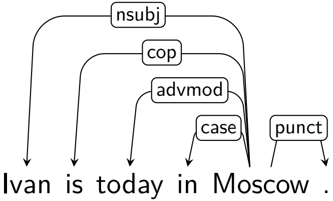


- This is also parallel with Russian:

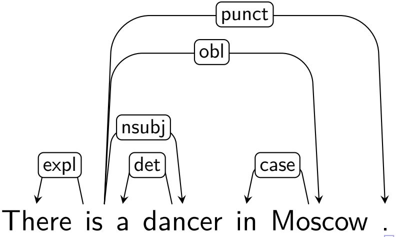


Well, Almost...

- This is parallel with Russian:

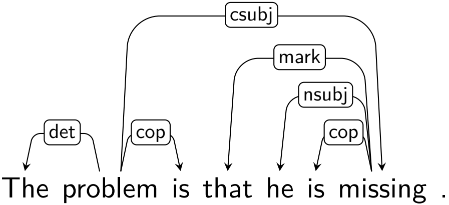


- But not with this in English:



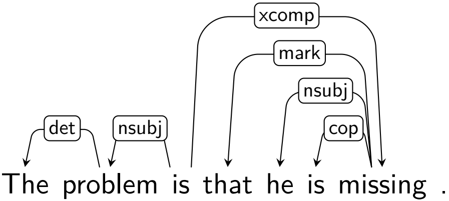
Clauses and Copula

- A clause can be the subject:



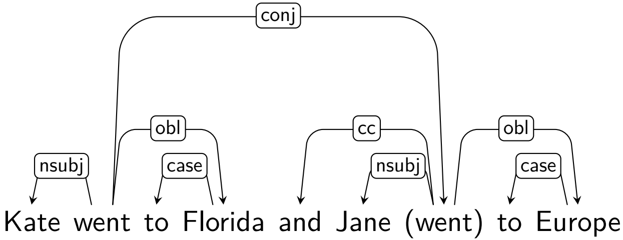
The problem is that he is missing .

- But it cannot be annotated as the nonverbal predicate:



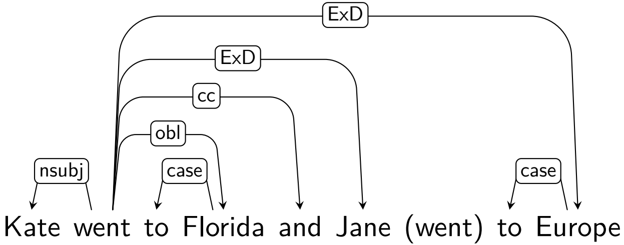
The problem is that he is missing .

Ellipsis: Deleted Predicates in Coordination

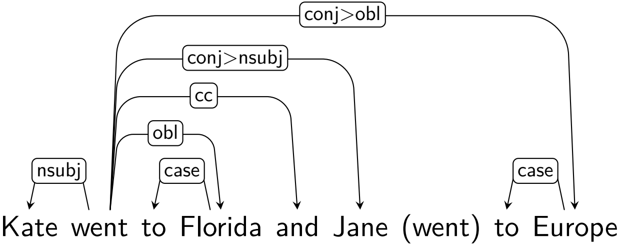


- Some treebanks would use an empty node to represent the second *went*.
- UD enhanced representation now allows empty nodes
- ... but the basic representation sticks with the overt words.

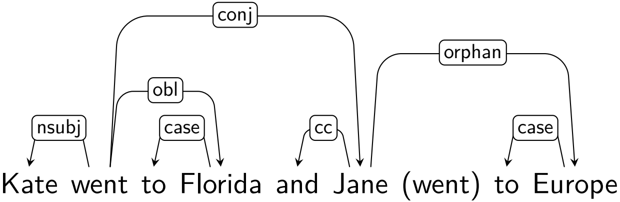
PDT: The ExD Relation



Perseus Treebanks: Chained Relations



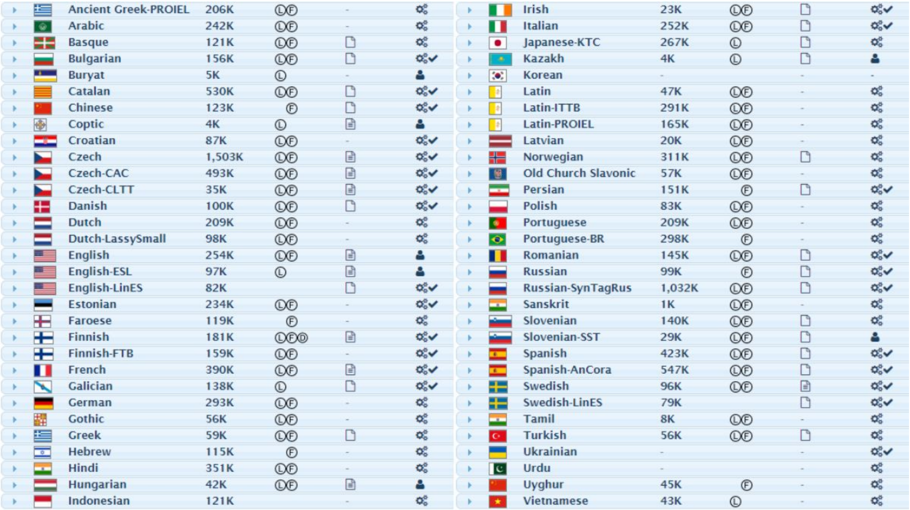
UD V2: The orphan Relation



Where Are We Now?

- Three years of UD
- 6 treebank releases (every 6 months)
- 95 treebanks, 57 languages (over 50% world's population)
- 11000+ unique IP downloads (all versions)
- Over 13M tokens; treebanks range from <1K to 1.5M
- Over 200 contributors
  - language group consistency SIGs
- Version 2 guidelines in place
- CoNLL Shared Task 2017 completed (ACL/CONLL) - coming soon

57 Languages and Growing



Ancient Greek-PROIEL	206K	ⓁⒻ	-	⚙
Arabic	242K	ⓁⒻ	-	⚙
Basque	121K	ⓁⒻ	🗋	⚙
Bulgarian	156K	ⓁⒻ	🗋	⚙✓
Buryat	5K	Ⓛ	-	👤
Catalan	530K	ⓁⒻ	🗋	⚙✓
Chinese	123K	Ⓕ	🗋	⚙✓
Coptic	4K	Ⓛ	🗎	👤
Croatian	87K	ⓁⒻ	-	⚙✓
Czech	1,503K	ⓁⒻ	🗎	⚙✓
Czech-CAC	493K	ⓁⒻ	🗎	⚙✓
Czech-CLTT	35K	ⓁⒻ	🗎	⚙✓
Danish	100K	ⓁⒻ	🗋	⚙✓
Dutch	209K	ⓁⒻ	-	⚙
Dutch-LassySmall	98K	ⓁⒻ	-	⚙
English	254K	ⓁⒻ	🗎	👤
English-ESL	97K	Ⓛ	🗎	👤
English-LinES	82K		🗋	⚙✓
Estonian	234K	ⓁⒻ	-	⚙✓
Faroese	119K	Ⓕ	-	⚙
Finnish	181K	ⓁⒻⒹ	🗎	⚙✓
Finnish-FTB	159K	ⓁⒻ	-	⚙✓
French	390K	ⓁⒻ	🗎	⚙✓
Galician	138K	Ⓛ	🗋	⚙✓
German	293K	ⓁⒻ	-	⚙
Gothic	56K	ⓁⒻ	-	⚙
Greek	59K	ⓁⒻ	🗋	⚙
Hebrew	115K	Ⓕ	-	⚙
Hindi	351K	ⓁⒻ	-	⚙
Hungarian	42K	ⓁⒻ	🗎	👤
Indonesian	121K		-	⚙
Irish	23K	ⓁⒻ	🗋	⚙✓
Italian	252K	ⓁⒻ	🗋	⚙✓
Japanese-KTC	267K	Ⓛ	🗋	⚙
Kazakh	4K	Ⓛ	🗋	👤
Korean	-		-	-
Latin	47K	ⓁⒻ	-	⚙
Latin-ITTB	291K	ⓁⒻ	-	⚙
Latin-PROIEL	165K	ⓁⒻ	-	⚙
Latvian	20K	ⓁⒻ	-	⚙
Norwegian	311K	ⓁⒻ	🗋	⚙
Old Church Slavonic	57K	ⓁⒻ	-	⚙
Persian	151K	Ⓕ	🗋	⚙✓
Polish	83K	ⓁⒻ	-	⚙
Portuguese	209K	ⓁⒻ	-	⚙
Portuguese-BR	298K	Ⓕ	-	⚙
Romanian	145K	ⓁⒻ	🗋	⚙✓
Russian	99K	Ⓕ	🗋	⚙✓
Russian-SynTagRus	1,032K	ⓁⒻ	🗋	⚙✓
Sanskrit	1K	ⓁⒻ	-	⚙
Slovenian	140K	ⓁⒻ	🗋	⚙
Slovenian-SST	29K	ⓁⒻ	🗋	👤
Spanish	423K	ⓁⒻ	🗋	⚙✓
Spanish-AnCora	547K	ⓁⒻ	🗋	⚙✓
Swedish	96K	ⓁⒻ	🗎	⚙✓
Swedish-LinES	79K		🗋	⚙✓
Tamil	8K	ⓁⒻ	-	⚙
Turkish	56K	ⓁⒻ	🗋	⚙
Ukrainian	-		-	⚙✓
Urdu	-		-	⚙
Uyghur	45K	Ⓕ	-	⚙
Vietnamese	43K	Ⓛ	-	⚙

Path to the CoNLL 2017 UD Shared Task

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)

Path to the CoNLL 2017 UD Shared Task

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)

Path to the CoNLL 2017 UD Shared Task

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)

Path to the CoNLL 2017 UD Shared Task

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
- SPMRL 2013 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- SPMRL 2014 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)

Path to the CoNLL 2017 UD Shared Task

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
- SPMRL 2013 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- SPMRL 2014 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- VarDial 2017 (cross-lingual: cs-sk, sl-hr, da/sv-no)

Path to the CoNLL 2017 UD Shared Task

- CoNLL 2006 (13 langs: ar, cs, bg, da, de, es, ja, nl, pt, sl, sv, tr, zh)
- CoNLL 2007 (10 langs: ar, ca, cs, el, en, eu, hu, it, tr, zh)
- CoNLL 2008: + semantic dependencies (English)
- CoNLL 2009: + semantic dependencies (ca, cs, de, en, es, ja, zh)
- ICON 2009 (Hindi, Bangla, Telugu)
- ICON 2010 (Hindi, Bangla, Telugu)
- SPMRL 2013 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- SPMRL 2014 (9 languages: ar, de, eu, fr, he, hu, ko, pl, sv)
- VarDial 2017 (cross-lingual: cs-sk, sl-hr, da/sv-no)
- CoNLL 2017 (45 languages + surprise + end-to-end parsing)

CoNLL 2017 UD ST Data: Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
  - Technical problem: Italian-ParTUT (overlap with Italian in test data)

CoNLL 2017 UD ST Data: Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
  - Technical problem: Italian-ParTUT (overlap with Italian in test data)
- Arabic NYUAD: not available free of charge

CoNLL 2017 UD ST Data: Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
  - Technical problem: Italian-ParTUT (overlap with Italian in test data)
- Arabic NYUAD: not available free of charge
- At least 10K test words ⇒
  - Exclude: Belarusian, Coptic, Lithuanian, Sanskrit, Tamil
  - Include but small training: French ParTUT, Galician TreeGal, Irish, **Kazakh**, Latin, Slovenian SST, Ukrainian, **Uyghur**

CoNLL 2017 UD ST Data: Languages and Treebanks

- All UD 2.0 treebanks except:
  - Too small
  - Non-free
  - Technical problem: Italian-ParTUT (overlap with Italian in test data)
- Arabic NYUAD: not available free of charge
- At least 10K test words ⇒
  - Exclude: Belarusian, Coptic, Lithuanian, Sanskrit, Tamil
  - Include but small training: French ParTUT, Galician TreeGal, Irish, **Kazakh**, Latin, Slovenian SST, Ukrainian, **Uyghur**
- Total of **63** treebanks in **45** languages

Additional Data

- Just one “closed” track
- Registered participants were asked for suggestions

- CommonCrawl + word embeddings
- Word Atlas of Language Structures (WALS)
- Wikipedia Dumps
  - Wikipedia word vectors (90 languages) by Facebook
- Opus Parallel Corpora
- WMT 2016 Parallel + Monolingual Data
- Apertium + Giellatekno Morphological Analyzers
- French Treebank UD v2 conversion

CoNLL 2017 UD Shared Task Evaluation Test Sets

- **81 test files in total**
- Evaluation test sets for “regular” UD languages with training data provided (63)
- Surprise languages (4)
  - Buryat, Kurdish, Northern Sámi, Upper Sorbian
- New parallel test sets (14, by DFKI, Google and others):
  - Task languages: sv tr pt ru it ja hi fr es fi en de cs ar
  - 4 others available now
- **Main system score:**
  - macro-average LAS across all test sets (not languages)
- A system must produce formally valid results on all 81 test sets to be counted in official results

End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data

End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data

- Sentence segmentation

End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data

- Sentence segmentation
- Tokenization
- Word segmentation (multi-word tokens)

End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data

- Sentence segmentation
- Tokenization
- Word segmentation (multi-word tokens)
- Morphological analysis
  - If your parser needs it
  - Exception: predicted morphology available for surprise languages

End-to-End Parsing

- A real-world scenario
- No gold-standard processing available in the test data

- Sentence segmentation
- Tokenization
- Word segmentation (multi-word tokens)
- Morphological analysis
  - If your parser needs it
  - Exception: predicted morphology available for surprise languages
- Parsing

Baseline Models

- UDPipe (ÚFAL): trained segmenter, tagger+lemmatizer, parser
- Pre-processed test data (except syntax) directly available
- Just use that if you don't have anything better

- SyntaxNet / ParseySaurus (Google)

- No interest in surprise languages?
  - Use simple delexicalized parser

Evaluation Metrics

- Align system-output tokens to gold tokens

Al-Zaman : American forces killed Shaikh Abdullah al-Ani, the preacher at the mosque in the town of Qaim, near the Syrian border.

GOLD:	**Al**	**-**	**Zaman**	**:**	**American**	**forces**	**killed**	**Shaikh**
OFFSET:	**0-1**	**2**	**3-7**	**9**	**11-18**	**20-25**	**27-32**	**34-39**

- All characters except for whitespace match => easy align!

SYSTEM:	**Al-Zaman**	**:**	**American**	**forces**	**killed**	**Shaikh**
OFFSET:	**0-7**	**9**	**11-18**	**20-25**	**27-32**	**34-39**

Evaluation Metrics

- Align system-output tokens to gold tokens

Die Kosten sind definitiv auch im Rahmen.

GOLD:	**Die**	**Kosten**	**sind**	**definitiv**	**auch**	**im**	**Rahmen**	**.**
SPLIT:	**Die**	**Kosten**	**sind**	**definitiv**	**auch**	**in dem**	**Rahmen**	**.**
OFFSET:	**0-2**	**4-9**	**11-14**	**16-24**	**26-29**	**31-32**	**34-39**	**40**

- Corresponding but not identical spans?
- Find longest common subsequence

SYSTEM:	**Kosten**	**sind**	**definitiv**	**auch**	**im**	**Rahmen**	**.**
SPLIT:	**Kosten**	**sind**	**de finitiv**	**auch**	**im**	**Rahmen**	**.**
OFFSET:	**4-9**	**11-14**	**16-24**	**26-29**	**31-32**	**34-39**	**40**

Evaluation Metrics

- Align system-output tokens to gold tokens

Die Kosten sind definitiv auch im Rahmen.

GOLD:	**Die**	**Kosten**	**sind**	**definitiv**	**auch**	**im**	**Rahmen**	**.**
SPLIT:	**Die**	**Kosten**	**sind**	**definitiv**	**auch**	**in dem**	**Rahmen**	**.**
OFFSET:	**0-2**	**4-9**	**11-14**	**16-24**	**26-29**	**31-32**	**34-39**	**40**

- Corresponding but not identical spans?
- Find longest common subsequence

SYSTEM:	**auch**	**im**	**Rahmen**	**.**
SPLIT:	**auch**	**in einem , dem alle zustimmen ,**	**Rahmen**	**.**
OFFSET:	**26-29**	**31-32**	**34-39**	**40**

Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
    - $head_{System}(i) = head_{Gold}(i)$
    - (Comparing just integers here.)

Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  - $head_{System}(i) = head_{Gold}(i)$
  - (Comparing just integers here.)
- Compare aligned nodes, if alignment is found
  - $node : Integer \to Node$
  - $align : SystemNode \to GoldNode$
  - $align(head_{System}(node_i)) = head_{Gold}(align(node_i))$
  - (Comparing node objects.)

Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  - $head_{System}(i) = head_{Gold}(i)$
  - (Comparing just integers here.)
- Compare aligned nodes, if alignment is found
  - $node : Integer \rightarrow Node$
  - $align : SystemNode \rightarrow GoldNode$
  - $align(head_{System}(node_i)) = head_{Gold}(align(node_i))$
  - (Comparing node objects.)
- **Cannot align? No point for attachment!**

Evaluation Metrics

- Word IDs no longer match between gold and system files!
- Instead of comparing gold HEAD to system HEAD
  - $head_{System}(i) = head_{Gold}(i)$
  - (Comparing just integers here.)
- Compare aligned nodes, if alignment is found
  - $node : Integer \rightarrow Node$
  - $align : SystemNode \rightarrow GoldNode$
  - $align(head_{System}(node_i)) = head_{Gold}(align(node_i))$
  - (Comparing node objects.)
- **Cannot align? No point for attachment!**
- Wrong sentence boundary?
  - one or more wrong relations

Main Evaluation Metrics: Labeled Attachment Score

- Point for “correct” relation:
  - alignment of parent equals to parent of alignment
  - universal prefix of dependency relation types matches on both sides

- Precision: $P = \frac{\#correctRelations}{\#systemNodes}$
- Recall: $R = \frac{\#correctRelations}{\#goldNodes}$

- LAS (labeled attachment F_1-score): $LAS = \frac{2PR}{P+R}$

- Average over 81 test files ⇒ main system score

Evaluation Style: Blind, on TIRA

- Strong recommendation of SIGNLL (new 2015):
- Teams submit software, not data
- TIRA evaluation platform
  - `http://www.tira.io/`

- Virtual machine for each team
  - Configurable number of CPUs, RAM, disk space
  - Currently no GPUs available
  - OS: Ubuntu, Fedora or Windows
  - Participants get admin access, can install anything
  - ⇒ **improved reproducibility**

Blind Evaluation on TIRA

- Running on test data:
  - Remote control through web interface (participants)
  - VM is “sandboxed”, detached from internet
- after the run:
  - Output files, STDOUT and STDERR archived in TIRA
  - State of VM before the run is restored (including disk)
  - Participants do not see any output
  - ⇒ **prevents test data leakage**

Blind Evaluation on TIRA

- Running on test data:
  - Remote control through web interface (participants)
  - VM is “sandboxed”, detached from internet
- after the run:
  - Output files, STDOUT and STDERR archived in TIRA
  - State of VM before the run is restored (including disk)
  - Participants do not see any output
  - ⇒ **prevents test data leakage**
  - **... but also makes the task extremely sensitive to mistakes**

#ParsingTragedy

- Debugging on development data (can see output)
  - but some files exist only in test data

#ParsingTragedy

- Debugging on development data (can see output)
  - but some files exist only in test data
- On-demand unblinding of runs by moderator

#ParsingTragedy

- Debugging on development data (can see output)
  - but some files exist only in test data
- On-demand unblinding of runs by moderator
- Cannot see scores on test data

#ParsingTragedy

- Debugging on development data (can see output)
  - but some files exist only in test data
- On-demand unblinding of runs by moderator
- Cannot see scores on test data
- System runs for two days
  - but nobody knows that it is stuck in an endless loop

#ParsingTragedy

- Debugging on development data (can see output)
    - but some files exist only in test data
- On-demand unblinding of runs by moderator
- Cannot see scores on test data

- System runs for two days
    - but nobody knows that it is stuck in an endless loop
    - or output files are not found
    - we had to stitch results from multiple runs

#ParsingTragedy

- Debugging on development data (can see output)
  - but some files exist only in test data
- On-demand unblinding of runs by moderator
- Cannot see scores on test data

- System runs for two days
  - but nobody knows that it is stuck in an endless loop
  - or output files are not found
  - we had to stitch results from multiple runs
- System finishes "successfully"
  - but when the results are announced you find out that it picked a wrong model

Participants

- 111 registrations

Participants

- 111 registrations
- 56 teams got virtual machine

Participants

- 111 registrations
- 56 teams got virtual machine
- 38 logged in the TIRA interface (plus 2 org. accounts, and 2 extra VMs)

Participants

- 111 registrations
- 56 teams got virtual machine
- 38 logged in the TIRA interface (plus 2 org. accounts, and 2 extra VMs)
- 34 ran something (plus 1 org. account: baseline)

Participants

- 111 registrations
- 56 teams got virtual machine
- 38 logged in the TIRA interface (plus 2 org. accounts, and 2 extra VMs)
- 34 ran something (plus 1 org. account: baseline)
- 32 reached non-zero score on test data

Participants

- 111 registrations
- 56 teams got virtual machine
- 38 logged in the TIRA interface (plus 2 org. accounts, and 2 extra VMs)
- 34 ran something (plus 1 org. account: baseline)
- 32 reached non-zero score on test data
- 27 reached non-zero on each of the 81 files

- (CoNLL 2006 had 17 participants)
- (CoNLL 2007 had 23 participants)

Results: Macro LAS F1

	Team	LAS	Files
1.	Stanford (Stanford)	76.30	[OK]
2.	C2L2 (Ithaca)	75.00	[OK]
3.	IMS (Stuttgart)	74.42	[OK]
4.	HIT-SCIR (Harbin)	72.11	[OK]
5.	LATTICE (Paris)	70.93	[OK]
6.	NAIST SATO (Nara)	70.14	[OK]
7.	Koç University (İstanbul)	69.76	[OK]
8.	ÚFAL – UDPipe 1.2 (Praha)	69.52	[OK]
9.	UParse (Edinburgh)	68.87	[OK]
10.	Orange – Deskiñ (Lannion)	68.61	[OK]
11.	TurkuNLP (Turku)	68.59	[OK]
12.	darc (Tübingen)	68.41	[OK]
13.	BASELINE UDPipe 1.1 (Praha)	68.35	[OK]

Unofficial Results #ParsingTragedy

	Team	LAS	Files
1.	Stanford (Stanford)	76.30	[OK]
2.	C2L2 (Ithaca)	75.00	[OK]
3.	IMS (Stuttgart)	74.42	[OK]
4.	HIT-SCIR (Harbin)	72.11	[OK]
5.	LATTICE (Paris)	70.93	[OK]
6.	ParisNLP (Paris)	70.35	[OK]
7.	NAIST SATO (Nara)	70.14	[OK]
8.	Koç University (İstanbul)	69.76	[OK]
9.	Uppsala (Uppsala)	69.66	[OK]
10.	ÚFAL – UDPipe 1.2 (Praha)	69.52	[OK]
11.	LyS-FASTPARSE (A Coruña)	69.15	[OK]
12.	LIMSI (Paris)	68.90	[OK]
13.	UParse (Edinburgh)	68.87	[OK]
14.	RACAI (București)	68.79	[OK]
15.	Orange – Deskiñ (Lannion)	68.63	[OK]

Results: Word Segmentation

	Team	**F$_1$**
1.	IMS (Stuttgart)	98.81
2.	LIMSI (Paris)	98.68
3.	ÚFAL – UDPipe 1.2 (Praha)	98.63
4.	HIT-SCIR (Harbin)	98.62
5.	ParisNLP (Paris)	98.58
6.	Wanghao-ftd-SJTU (Shanghai)	98.55
	darc (Tübingen)	98.55
8.	BASELINE UDPipe 1.1 (Praha)	98.50
	C2L2 (Ithaca)	98.50
	IIT Kharagpur (Kharagpur)	98.50
	Koç University (İstanbul)	98.50
	LATTICE (Paris)	98.50
	LyS-FASTPARSE (A Coruña)	98.50
	METU (Ankara)	98.50
	MQuni (Sydney)	98.50

CLAS: a UD-specific Weighted Metric (Experimental)

- Relations between content words are more important cross-linguistically
- Attachment of function word = morphology in other languages
- Weighted scoring of correct relations:
  - **Weight = 1** for *root, nsubj, obj, iobj, csubj, ccomp, xcomp, obl, vocative, expl, dislocated, advcl, advmod, discourse, nmod, appos, nummod, acl, amod, conj, fixed, flat, compound, list, parataxis, orphan, goeswith, reparandum, dep*
  - **Weight = 0** for *aux, case, cc, clf, cop, det, mark*
  - **Weight = 0** for *punct*

Results: Macro CLAS

	Team	CLAS F_1	LAS F_1
1.	Stanford (Stanford)	72.57	76.30
2.	C2L2 (Ithaca)	70.91	75.00
3.	IMS (Stuttgart)	70.18	74.42
4.	HIT-SCIR (Harbin)	67.63	72.11
5.	LATTICE (Paris)	66.16	70.93
6.	NAIST SATO (Nara)	65.15	70.14
7.	Koç University (İstanbul)	64.61	69.76
8.	ÚFAL – UDPipe 1.2 (Praha)	64.36	69.52
9.	Orange – Deskiñ (Lannion)	64.15	68.61
10.	TurkuNLP (Turku)	63.61	68.59
11.	UParse (Edinburgh) (was: 9)	63.55	68.87
12.	darc (Tübingen)	63.24	68.41
13.	BASELINE UDPipe 1.1 (Praha)	63.02	68.35

Results: Surprise Languages

	Team	LAS F_1
1.	C2L2 (Ithaca)	47.54
2.	IMS (Stuttgart)	45.32
3.	HIT-SCIR (Harbin)	42.64
4.	Stanford (Stanford)	40.57
5.	ParisNLP (Paris)	39.23
6.	UParse (Edinburgh)	39.17
7.	Koç University (İstanbul)	38.81
8.	Orange – Deskiñ (Lannion)	38.72
9.	LIMSI (Paris)	37.57
10.	IIT Kharagpur (Kharagpur)	37.17
11.	BASELINE UDPipe 1.1 (Praha)	37.07

Results: Treebank Ranking by LAS

	Treebank	Max	MaxTeam	Avg	StDev
1.	ru_syntagrus	92.60	Stanford	71.64	±15.20
2.	hi	91.59	Stanford	73.41	±25.06
3.	sl	91.51	Stanford	69.70	±23.96
4.	pt_br	91.36	Stanford	72.58	±21.58
5.	ja	91.13	TRL	64.99	±23.45
6.	ca	90.70	Stanford	73.55	±21.10
7.	it	90.68	Stanford	74.06	±21.09
8.	cs_cac	90.43	Stanford	71.20	±12.07
9.	pl	90.32	Stanford	69.11	±21.59
10.	cs	90.17	Stanford	69.62	±12.34
11.	es_ancora	89.99	Stanford	72.53	±11.16
12.	no_bokmaal	89.88	Stanford	70.73	±20.97
13.	bg	89.81	Stanford	74.40	±20.46
14.	no_nynorsk	88.81	Stanford	66.81	±23.54
15.	fi_pud	88.47	Stanford	62.75	±19.28

Results: Treebank Ranking by CLAS

	Treebank	Max	MaxTeam	Avg	StDev
1.	ru_syntagrus	90.11	Stanford	67.83	±14.94
2.	sl	88.98	Stanford	65.77	±23.26
3.	cs	88.44	Stanford	66.98	±12.27
4.	cs_cac	88.31	Stanford	67.92	±11.89
5.	pl	87.94	Stanford	65.30	±20.61
6.	hi	87.92	Stanford	68.23	±24.29
7.	no_bokmaal	87.67	Stanford	67.18	±20.55
8.	pt_br	87.48	Stanford	66.36	±21.42
9.	fi_pud	86.82	Stanford	60.88	±18.25
10.	ca	86.70	Stanford	67.55	±20.36
11.	bg	86.53	Stanford	69.61	±20.13
12.	no_nynorsk	86.41	Stanford	62.92	±22.96
13.	it	86.18	Stanford	68.18	±19.79
14.	es_ancora	86.15	Stanford	66.90	±11.73
15.	nl_lassysmall	85.22	Stanford	63.61	±22.73

Thank You!
Questions?

`http://universaldependencies.org/`

`http://universaldependencies.org/conll17/`

UD Official repository: `http://lindat.cz/`