UniDive Training School

Funded by the European Union

Session 14

Error mining

Bruno Guillaume

This is joint work with **Guillaume Bonfante**, **Guy Perrier**, **Kim Gerdes**, **Sylvain Kahane**, **Gaël Guibon**, **Kirian Guiller**, **Khensa Daoudi** and others

Error mining in UD

UD_English-LinES@2.14

observe occurrences of `cc`

```
pattern { X -[cc]-> Y }
```

⚙

(1) have a `CCONJ` as dependent
(2) are right-headed

Exceptions to (1)?

```
pattern { X -[cc]-> Y; Y [upos <> CCONJ] }
```

⚙ ✅

Exceptions to (2)?

```
pattern { X -[cc]-> Y; X << Y }
```

⚙ ❌ 8 annotations to be checked

Error mining

number agreement with `subj`

observe occurrences of `subj`
without number agreement

```
pattern {
  V -[nsubj]-> S;
  V.Number <> S.Number
}
```

Explore and error mining: relation tables

- On each treebank, a set of **relations tables** (one per relation) is available
- Equivalent* to a double clustering of `upos` of the governor / `upos` of the dependent

* `ExtPos` is taken into account if present on the dependent

Go to `UD_English-LinES@2.14`

Use: ▦ and chose `amod` relation

- In ArboratorGrew, tables are available with the bottom right button

Go to `UD_Italian-PUD`

Use: ▦ and chose `nsubj` relation

Lexicon in ArboratorGrew

See documentation

- A "lexicon" can be computed on the current data
  - select a tuple **T** of features* ➔ AG computes the set of possible tuple of values for **T**. If inconsistencies are found, they can be fixed easily
    - Ex: on `SUD_Zaar-Autogramm`: `(Mood)`
    - Ex: on `SUD_Zaar-Autogramm`: `(Mood, upos)`
  - select two disjoint tuples **T** and **U** of features* ➔ AG computes the set of possible values such that for the values associated to **T** are associated with more than one set of values associated to **U** (**T** is ambiguous wrt **U**)
    - Ex: on `UD_Italian-PUD`, find all values of a `(form,lemma)` for which `upos` annotation is ambiguous
    - Ex: on `UD_Italian-PUD`, find all values of a `(form,lemma,upos)` for which `Gender` annotation is ambiguous

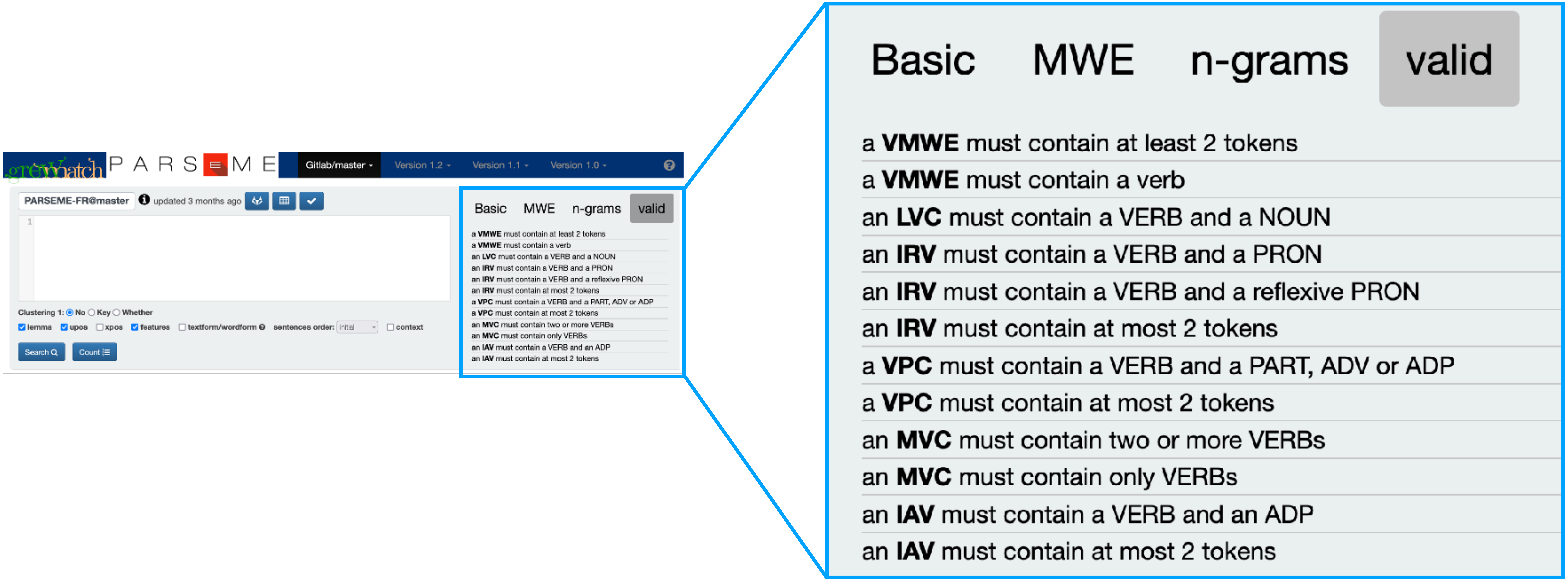
Error mining in Parseme

```
pattern {MWE [label]}
without {MWE -> V; V[upos=VERB]}
```

```
pattern {MWE [label = IRV]}
without {
  MWE -> V; V[upos=VERB];
  MWE -> P; P[upos=PRON, Reflex=Yes]
}
```

Error mining: consistency with UD

Many other examples available in the online interface



Clustering 1: ● No ○ Key ○ Whether

☑ lemma ☑ upos □ xpos ☑ features □ textform/wordform sentences order: initial □ context

a **VMWE** must contain at least 2 tokens

a **VMWE** must contain a verb

an **LVC** must contain a VERB and a NOUN

an **IRV** must contain a VERB and a PRON

an **IRV** must contain a VERB and a reflexive PRON

an **IRV** must contain at most 2 tokens

a **VPC** must contain a VERB and a PART, ADV or ADP

a **VPC** must contain at most 2 tokens

an **MVC** must contain two or more VERBs

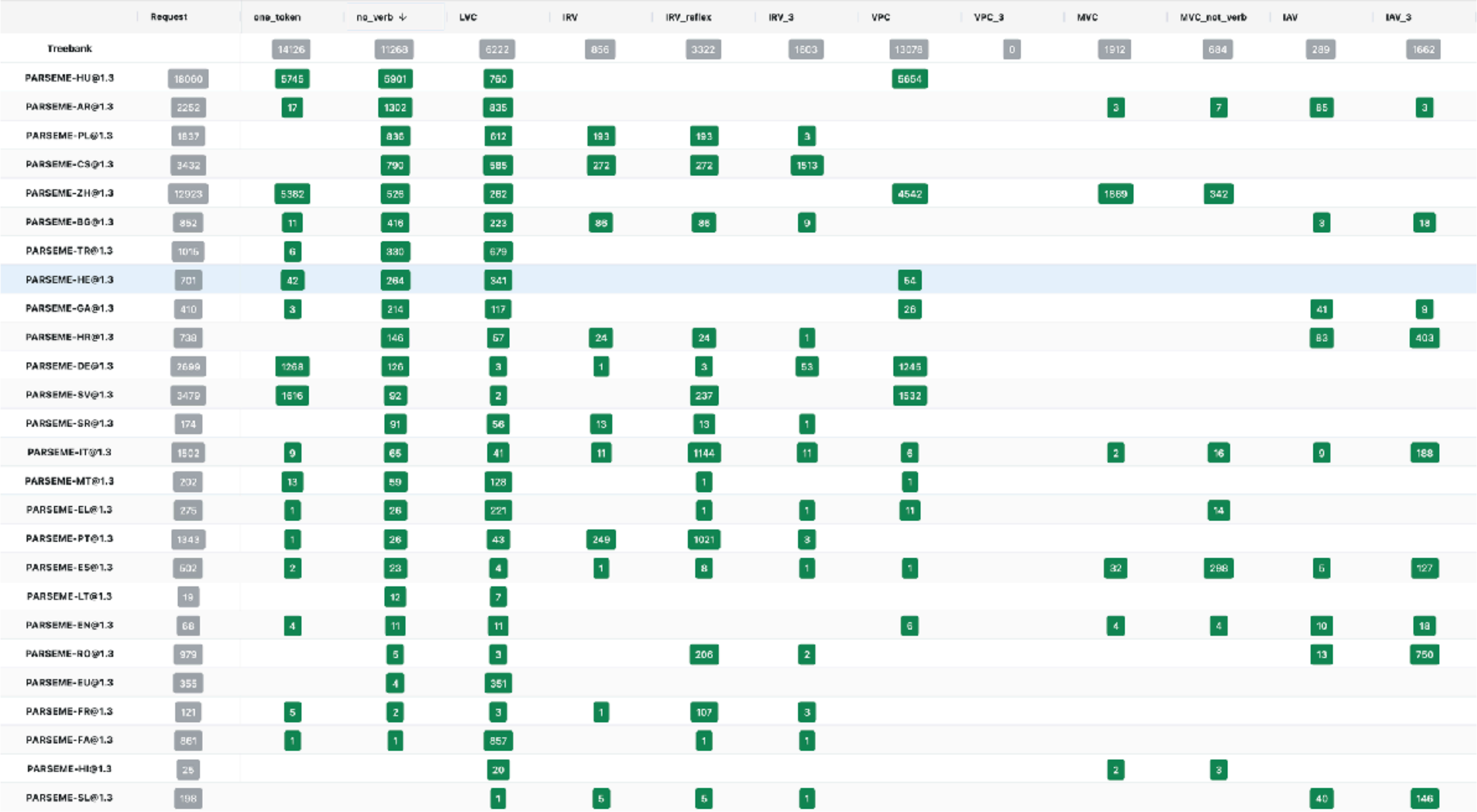
an **MVC** must contain only VERBs

an **IAV** must contain a VERB and an ADP

an **IAV** must contain at most 2 tokens

http://parseme.grew.fr

Error mining: consistency with UD



	Request	one_token	no_verb ↓	LVC	IRV	IRV_reflex	IRV_3	VPC	VPC_3	MVC	MVC_not_verb	IAV	IAV_3
Treebank		14126	11263	6222	856	3322	1603	13078	0	1912	684	289	1662
PARSEME-HU@1.3	16060	5745	5901	760				5654					
PARSEME-AR@1.3	2252	17	1302	835						3	7	65	3
PARSEME-PL@1.3	1637		836	612	193	193	3						
PARSEME-CS@1.3	3432		793	585	272	272	1513						
PARSEME-ZH@1.3	12923	5382	526	262				4542		1669	342		
PARSEME-BG@1.3	852	11	416	223	88	88	9					3	18
PARSEME-TR@1.3	1016	6	330	679									
PARSEME-HE@1.3	701	42	264	341				54					
PARSEME-GA@1.3	410	3	214	117				26				41	9
PARSEME-HR@1.3	738		146	67	24	24	1					83	403
PARSEME-DE@1.3	2699	1268	126	3	1	3	53	1245					
PARSEME-SV@1.3	3479	1616	92	2		237		1532					
PARSEME-SR@1.3	174		91	56	13	13	1						
PARSEME-IT@1.3	1502	9	65	41	11	1144	11	6		2	16	9	188
PARSEME-MT@1.3	202	13	59	128		1		1					
PARSEME-EL@1.3	275	1	26	221		1	1	11			14		
PARSEME-PT@1.3	1343	1	26	43	249	1021	3						
PARSEME-ES@1.3	602	2	23	4	1	8	1	1		32	298	6	127
PARSEME-LT@1.3	19		12	7									
PARSEME-EN@1.3	68	4	11	11				6		4	4	10	18
PARSEME-RO@1.3	979		5	3		206	2					13	750
PARSEME-EU@1.3	355		4	351									
PARSEME-FR@1.3	121	5	2	3	1	107	3						
PARSEME-FA@1.3	861	1	1	857		1	1						
PARSEME-HI@1.3	25			20						2	3		
PARSEME-SL@1.3	198			1	5	5	1					40	146

https://parseme.grew.fr/tables/?data=parseme/valid@1.3