Graph-based multi-layer querying in Parseme Corpora

Bruno Guillaume — LORIA / Inria Nancy Grand-Est

19th Workshop on Multiword Expressions (MWE 2023)

Dubrovnik / Online

PARSEME

- Annotation project for **Verbal Multi-Word Expressions**
- Available in **26 languages** (release 1.3, 2023)
- All Parseme corpora are released with **UD annotations** of the sentences
  - Annotation of Parseme on the top of **UD data**
  - Automatic parsing with **UDPIPE**

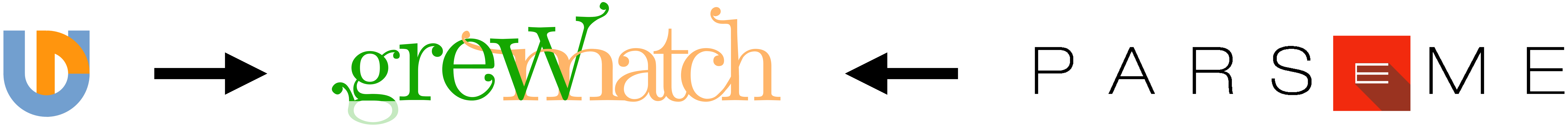
https://gitlab.com/parseme/corpora

grewmatch

- Web interface for online **requests** on annotated corpora
- Based on **graph representation** of the linguistic data
- **Available** on **syntactic** treebanks (UD, SUD…), on **semantic** graphs

http://match.grew.fr

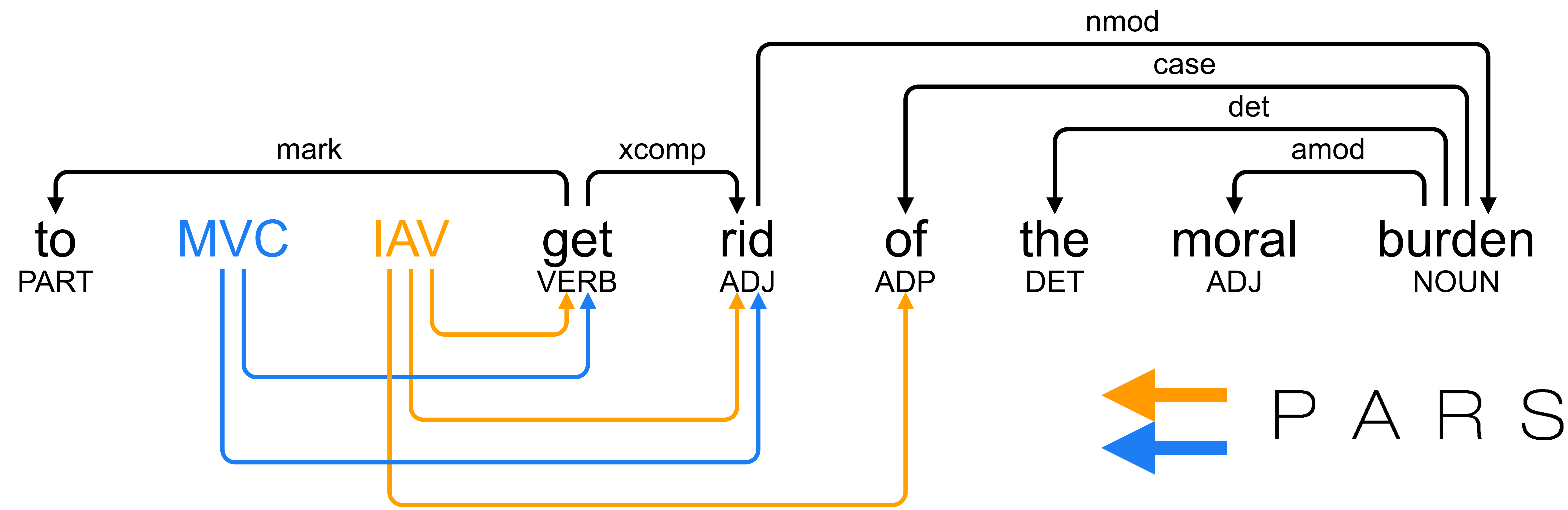
Multi-layer querying



- Encode the **two annotation layers** in the **same** graph structure
  - Parseme **VMWEs** can be **discontinuous**
  - Several Parseme **VMWEs** can **overlap**

- Each **VMWE** is a new **node**
- **Edges** map each **VMWE** node to all **its tokens**

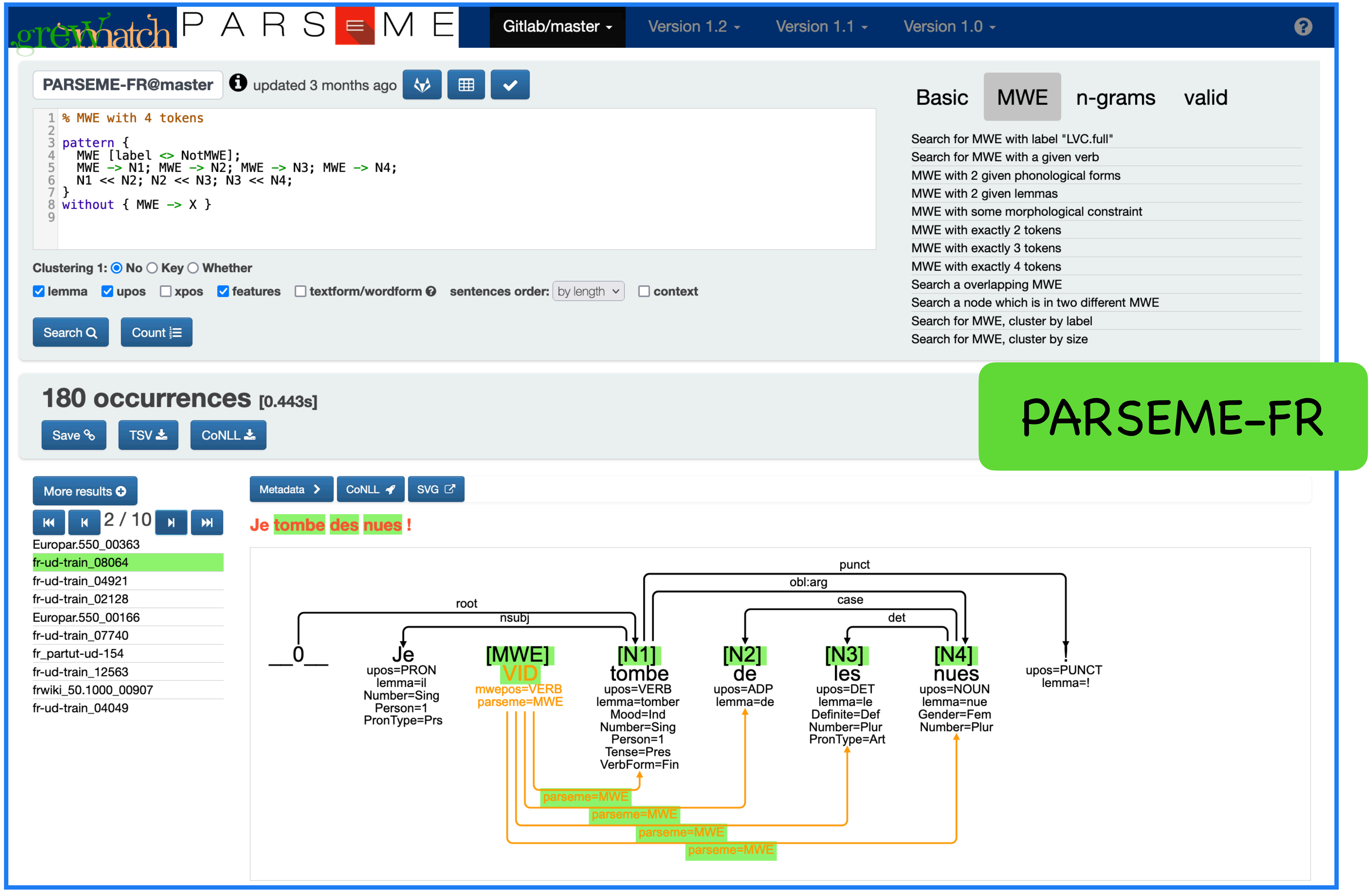
PARSEME-EN



PARSEME

grewmatch

- **Online** web queries
- **26 languages**
- Specific **query language**
- **Tutorial** available



grewmatch PARSEME | Gitlab/master ▾ | Version 1.2 ▾ | Version 1.1 ▾ | Version 1.0 ▾ | ?

PARSEME-FR@master ℹ updated 3 months ago

```
% MWE with 4 tokens

pattern {
  MWE [label <> NotMWE];
  MWE -> N1; MWE -> N2; MWE -> N3; MWE -> N4;
  N1 << N2; N2 << N3; N3 << N4;
}
without { MWE -> X }
```

Clustering 1: ● No ○ Key ○ Whether

☑ lemma ☑ upos □ xpos ☑ features □ textform/wordform ? sentences order: by length □ context

Search | Count

Basic | MWE | n-grams | valid

- Search for MWE with label "LVC.full"
- Search for MWE with a given verb
- MWE with 2 given phonological forms
- MWE with 2 given lemmas
- MWE with some morphological constraint
- MWE with exactly 2 tokens
- MWE with exactly 3 tokens
- MWE with exactly 4 tokens
- Search a overlapping MWE
- Search a node which is in two different MWE
- Search for MWE, cluster by label
- Search for MWE, cluster by size

180 occurrences [0.443s]

Save | TSV | CoNLL

More results

2 / 10

- Europar.550_00363
- fr-ud-train_08064
- fr-ud-train_04921
- fr-ud-train_02128
- Europar.550_00166
- fr-ud-train_07740
- fr_partut-ud-154
- fr-ud-train_12563
- frwiki_50.1000_00907
- fr-ud-train_04049

Metadata > | CoNLL | SVG

Je tombe des nues !

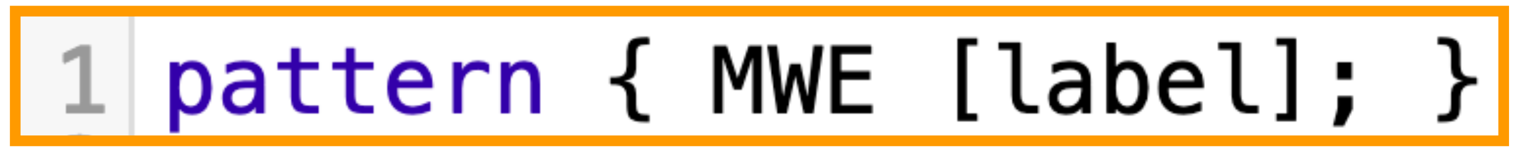
PARSEME-FR

http://parseme.grew.fr

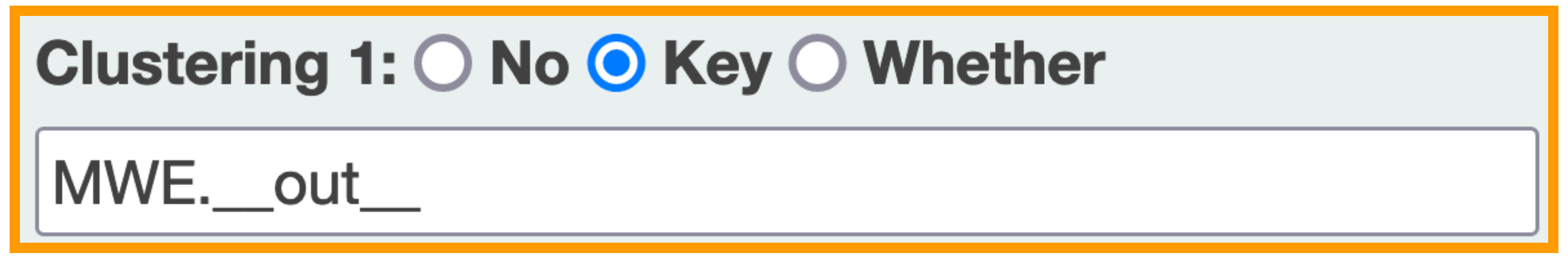
Je	*tombe*	*des*		*nues*	*!*
I	*fall*	*off*	*the*	*clouds (old form)*	*!*

`I can't believe it!`

VMWEs by size



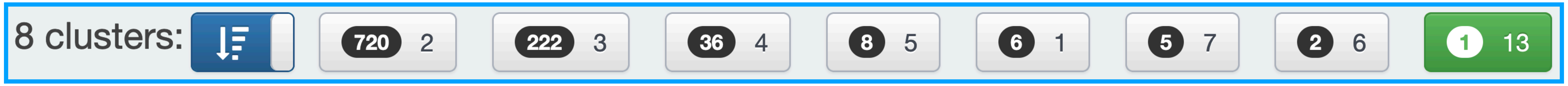
```
pattern { MWE [label]; }
```



Clustering 1: ○ No ● Key ○ Whether

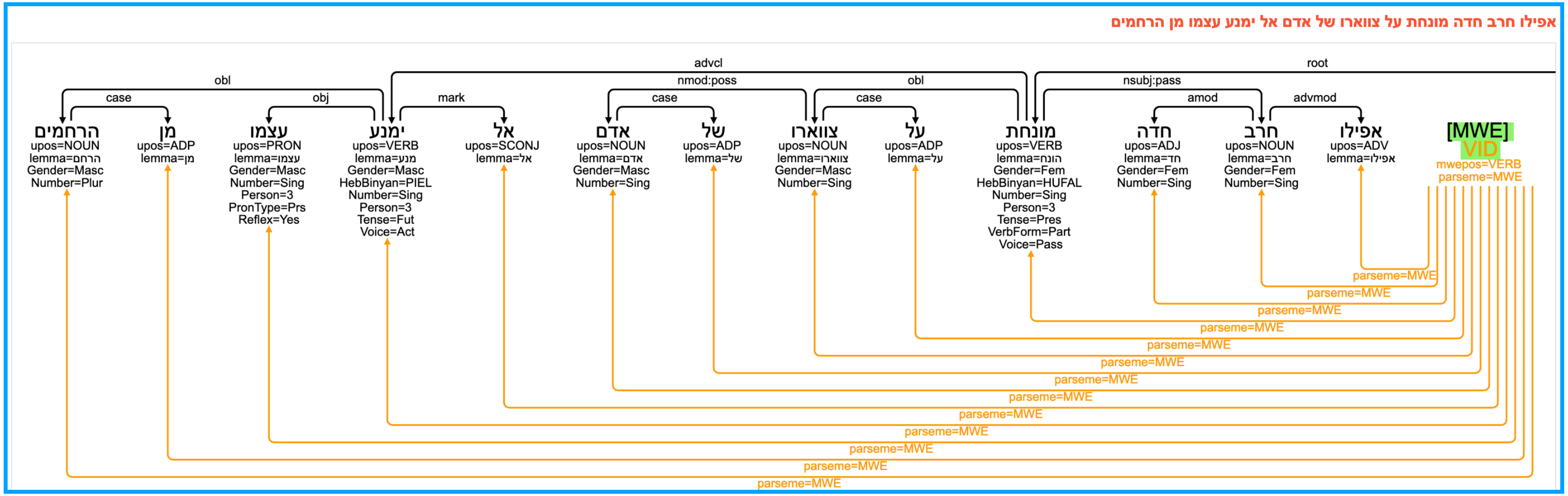
MWE.__out__

PARSEME-HE



8 clusters:

720	2
222	3
36	4
8	5
6	1
5	7
2	6
1	13

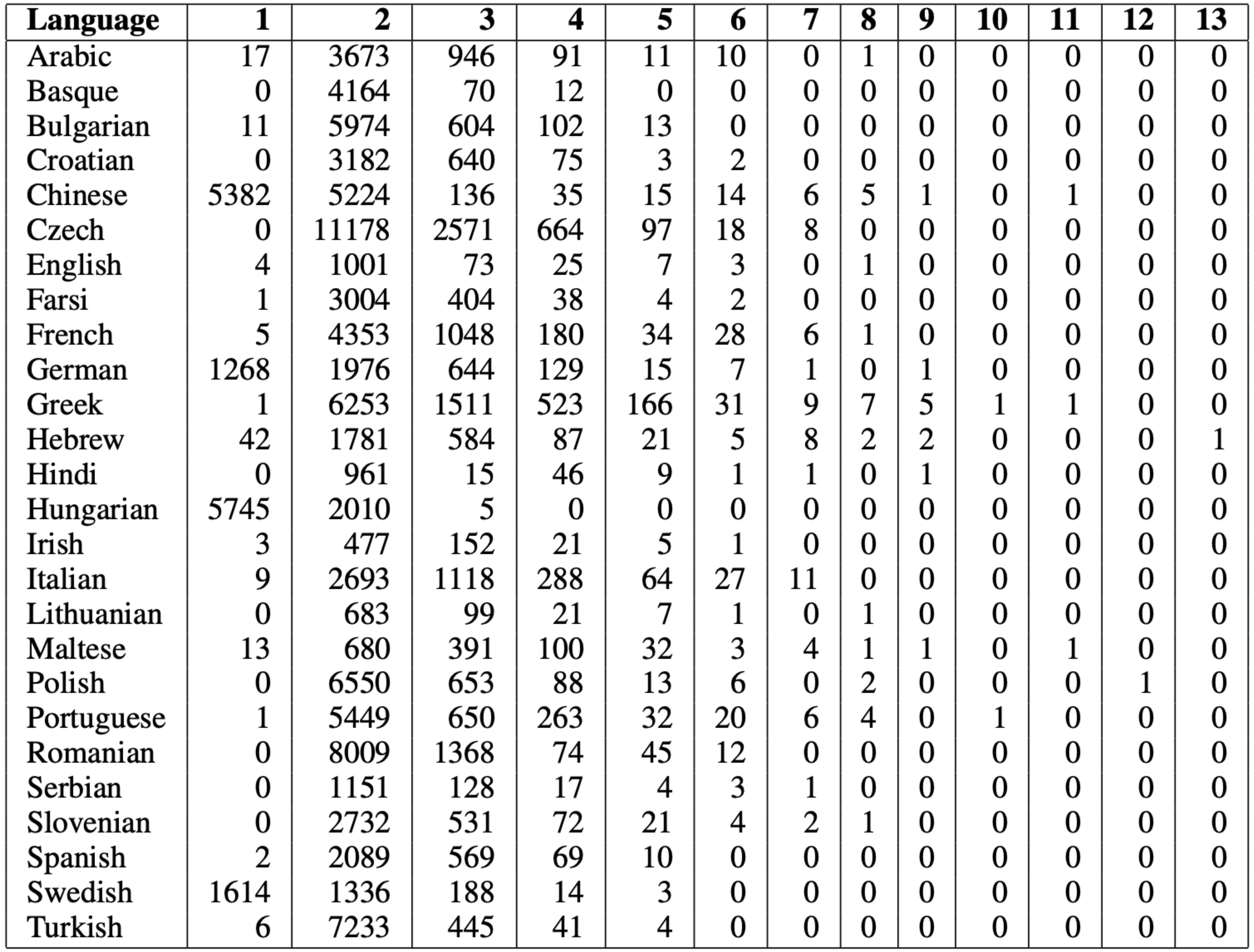


אפילו חרב חדה מונחת על צווארו של אדם אל ימנע עצמו מן הרחמים

root
advcl
obl
nmod:poss
obl
nsubj:pass
case
obj
mark
case
case
amod
advmod
הרחמים upos=NOUN lemma=הרחם Gender=Masc Number=Plur
מן upos=ADP lemma=מן
עצמו upos=PRON lemma=עצמו Gender=Masc Number=Sing Person=3 PronType=Prs Reflex=Yes
ימנע upos=VERB lemma=מנע Gender=Masc HebBinyan=PIEL Number=Sing Person=3 Tense=Fut Voice=Act
אל upos=SCONJ lemma=אל
אדם upos=NOUN lemma=אדם Gender=Masc Number=Sing
של upos=ADP lemma=של
צווארו upos=NOUN lemma=צוואר Gender=Masc Number=Sing
על upos=ADP lemma=על
מונחת upos=VERB lemma=הונח Gender=Fem HebBinyan=HUFAL Number=Sing Person=3 Tense=Pres VerbForm=Part Voice=Pass
חדה upos=ADJ lemma=חד Gender=Fem Number=Sing
חרב upos=NOUN lemma=חרב Gender=Fem Number=Sing
אפילו upos=ADV lemma=אפילו
[MWE] VID mwepos=VERB parseme=MWE
parseme=MWE

VMWEs by size

- Requests can be used in Python scripts (**grewpy**)
- **Ex:** run the requests about size on all the Parseme treebanks

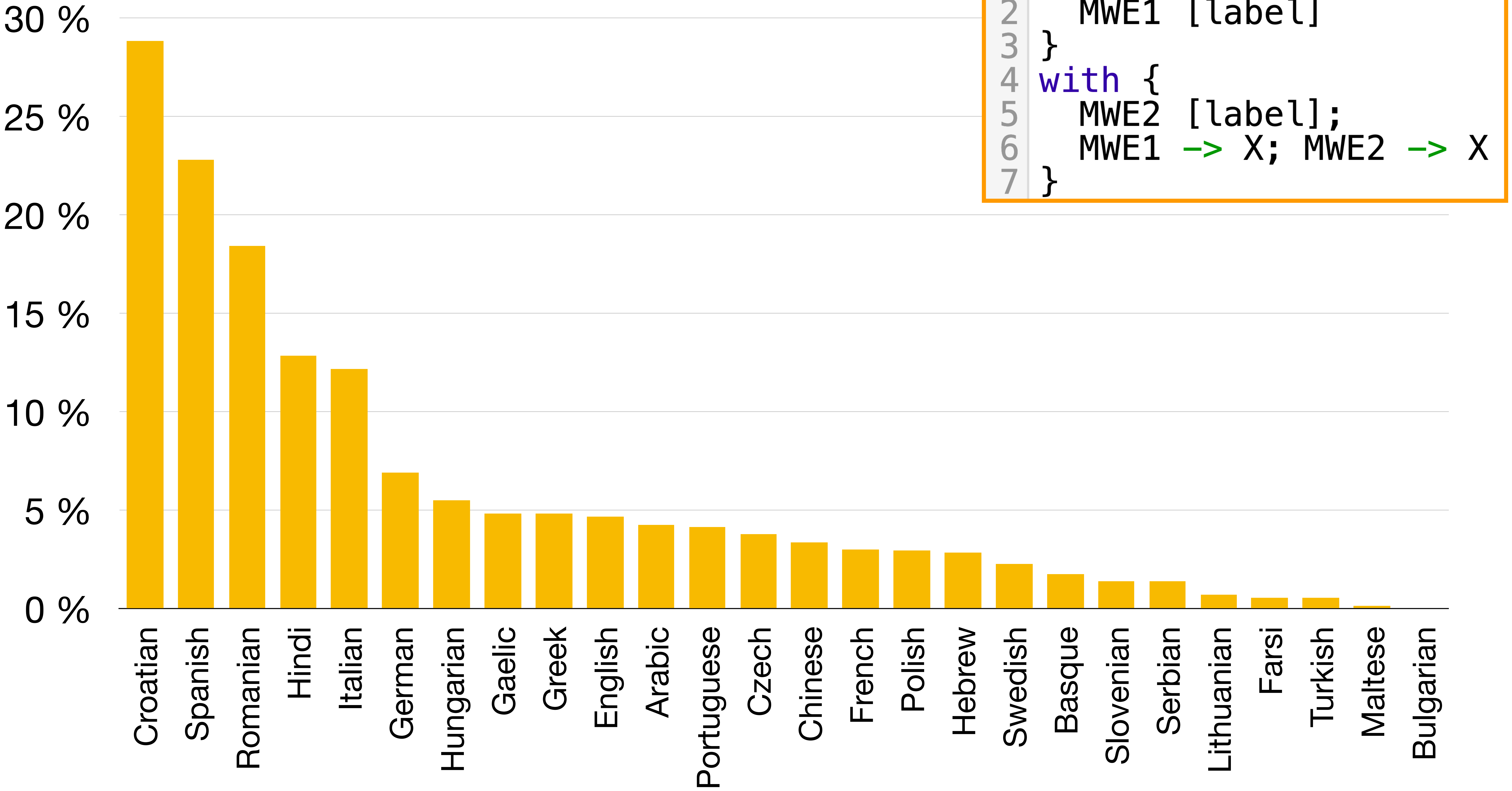


Language	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Arabic	17	3673	946	91	11	10	0	1	0	0	0	0	0
Basque	0	4164	70	12	0	0	0	0	0	0	0	0	0
Bulgarian	11	5974	604	102	13	0	0	0	0	0	0	0	0
Croatian	0	3182	640	75	3	2	0	0	0	0	0	0	0
Chinese	5382	5224	136	35	15	14	6	5	1	0	1	0	0
Czech	0	11178	2571	664	97	18	8	0	0	0	0	0	0
English	4	1001	73	25	7	3	0	1	0	0	0	0	0
Farsi	1	3004	404	38	4	2	0	0	0	0	0	0	0
French	5	4353	1048	180	34	28	6	1	0	0	0	0	0
German	1268	1976	644	129	15	7	1	0	1	0	0	0	0
Greek	1	6253	1511	523	166	31	9	7	5	1	1	0	0
Hebrew	42	1781	584	87	21	5	8	2	2	0	0	0	1
Hindi	0	961	15	46	9	1	1	0	1	0	0	0	0
Hungarian	5745	2010	5	0	0	0	0	0	0	0	0	0	0
Irish	3	477	152	21	5	1	0	0	0	0	0	0	0
Italian	9	2693	1118	288	64	27	11	0	0	0	0	0	0
Lithuanian	0	683	99	21	7	1	0	1	0	0	0	0	0
Maltese	13	680	391	100	32	3	4	1	1	0	1	0	0
Polish	0	6550	653	88	13	6	0	2	0	0	0	1	0
Portuguese	1	5449	650	263	32	20	6	4	0	1	0	0	0
Romanian	0	8009	1368	74	45	12	0	0	0	0	0	0	0
Serbian	0	1151	128	17	4	3	1	0	0	0	0	0	0
Slovenian	0	2732	531	72	21	4	2	1	0	0	0	0	0
Spanish	2	2089	569	69	10	0	0	0	0	0	0	0	0
Swedish	1614	1336	188	14	3	0	0	0	0	0	0	0	0
Turkish	6	7233	445	41	4	0	0	0	0	0	0	0	0

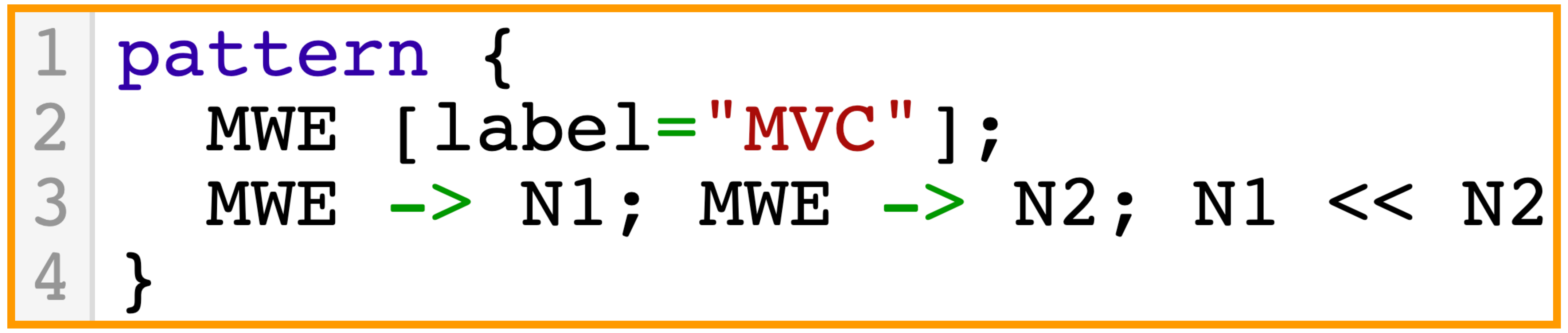
https://grew.fr/usage/python/

Overlapping VMWEs

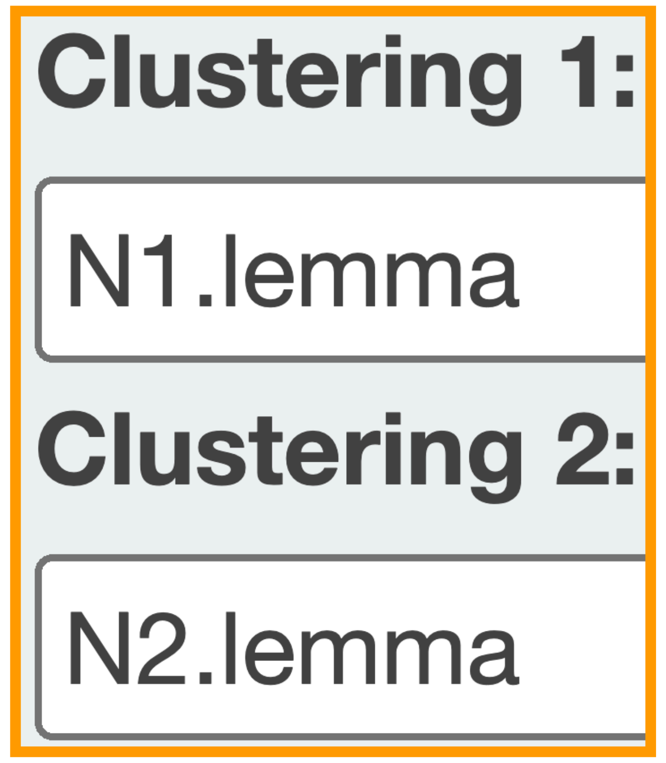
```
pattern {
  MWE1 [label]
}
with {
  MWE2 [label];
  MWE1 -> X; MWE2 -> X
}
```



Lemmas used in MVC annotations



```
pattern {
  MWE [label="MVC"];
  MWE -> N1; MWE -> N2; N1 << N2
}
```



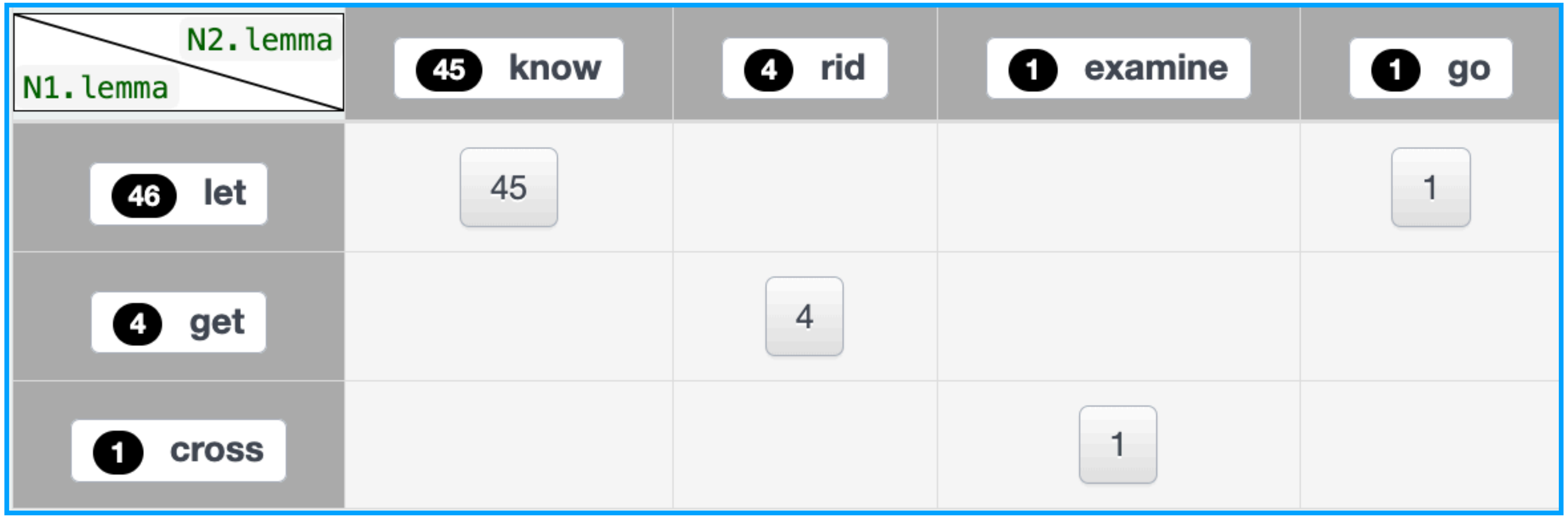
Clustering 1:

N1.lemma

Clustering 2:

N2.lemma

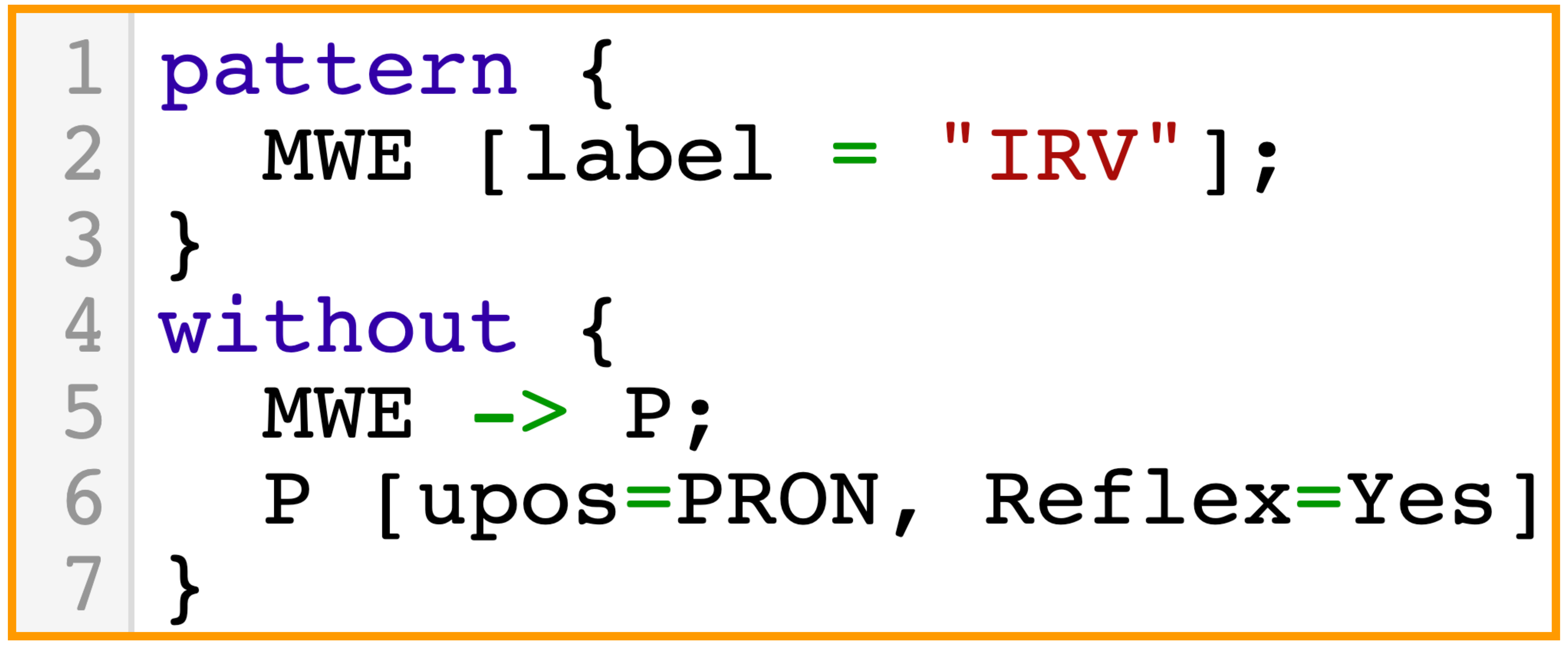
PARSEME-EN



N2.lemma N1.lemma	45 know	4 rid	1 examine	1 go
46 let	45			1
4 get		4		
1 cross			1	

Error mining: consistency with UD

Example: an **IRV** without a **reflexive pronoun**?

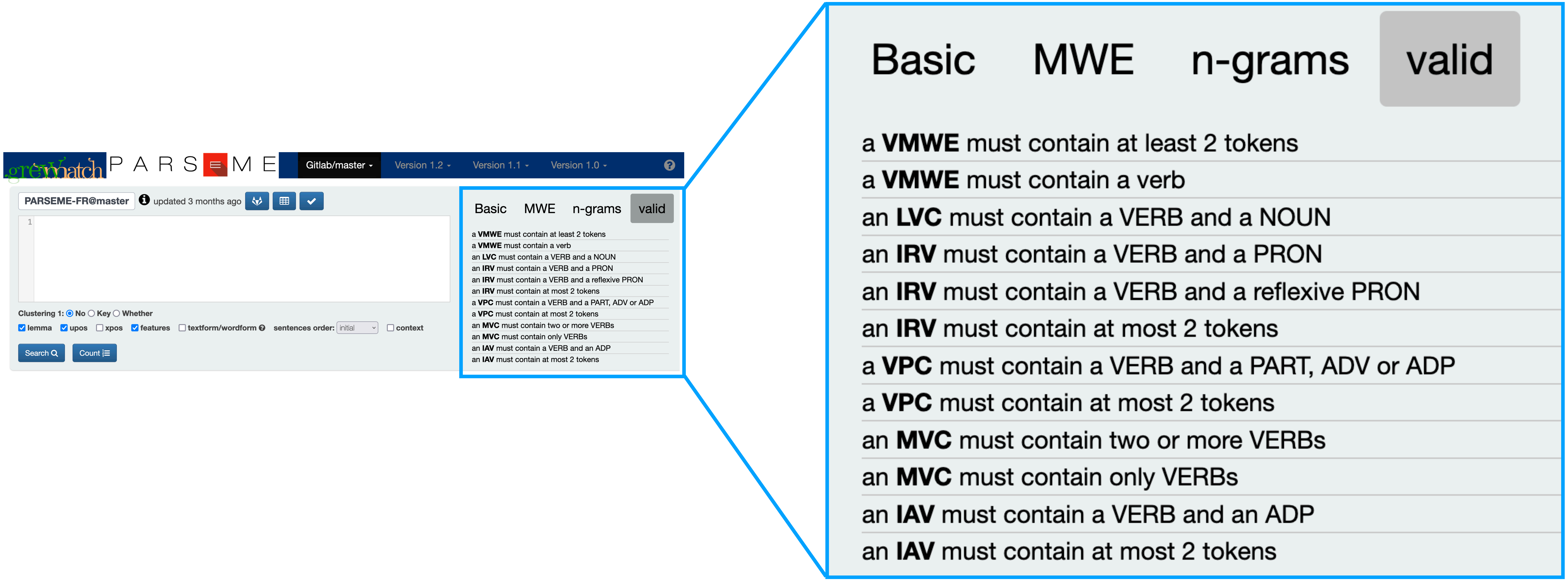


```
pattern {
  MWE [label = "IRV"];
}
without {
  MWE -> P;
  P [upos=PRON, Reflex=Yes]
}
```

	IRV without Reflex PRON	Reflex PRON	IRV without PRON
PARSEME-IT	1144	0	8
PARSEME-PT	1021	0	249
PARSEME-SV	237	0	0
PARSEME-RO	206	8863	0
PARSEME-FR	107	2806	1
PARSEME-ES	8	2120	1

Error mining: consistency with UD

Many other examples available in the online interface



PARSEME-FR@master updated 3 months ago

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

Conclusion

- **Graphs** can be used as a efficient way of connecting different **annotation layers**
- **Grew** implements graph-based structures for **NLP**
    - **Pattern Graph matching** in **Grew-match** (linguistic observations and error mining)
    - **Graph Rewriting** in **Grew** (conversion, consistent updates)
- Several **interfaces**
    - **Grew-match**
    - Python library: **grewpy**
    - Grew Command Line Interface
    - **Grew-web**: online rewriting, for testing and debugging

grew

`https://grew.fr`