cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

UniDive Training School

Funded by
the European Union

Session 13

Advanced treebank querying and annotation

Bruno Guillaume

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

Outline

- Using **Grew-match** on **Parseme** data

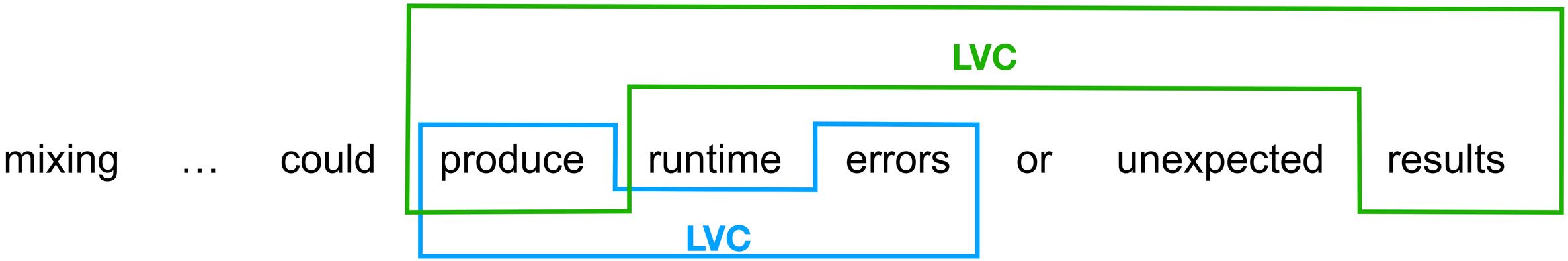
- Corpus pre-annotation with **ArboratorGrew**

Parseme data

Main portal: https://parseme.grew.fr

- 27 "latest" treebanks (26 from Parseme + Dutch in preparation)
- 26 from Parseme version 1.3
- 14 from version 1.2

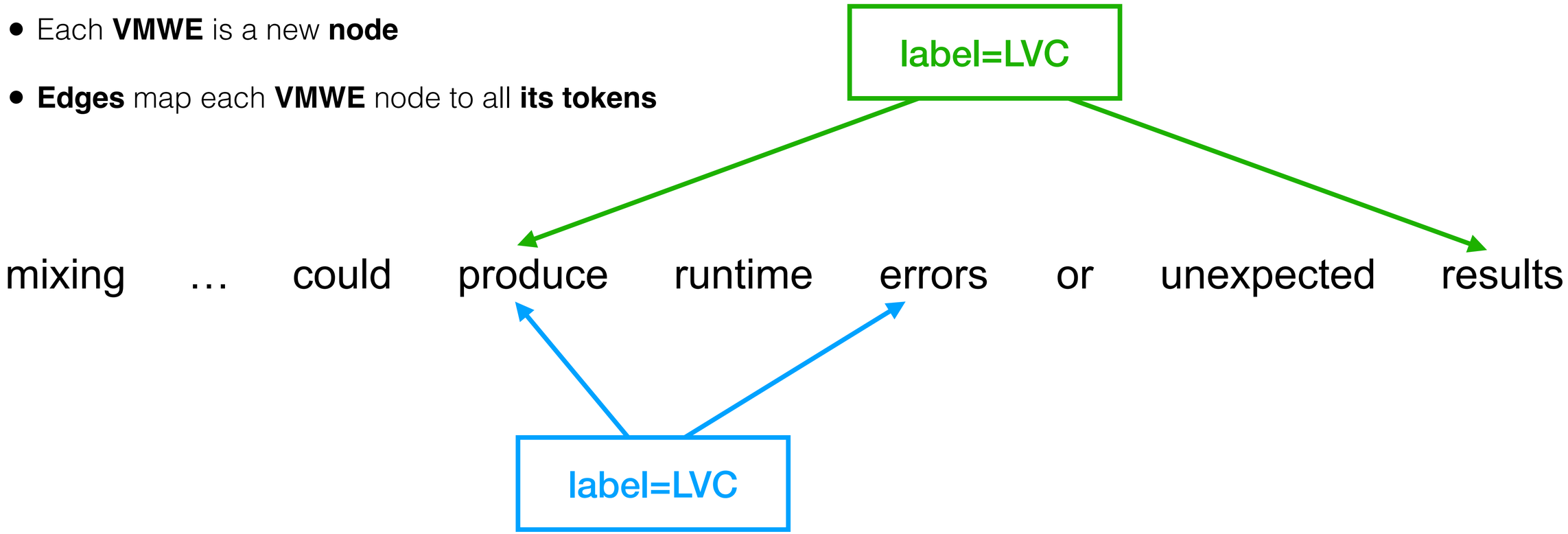
grewmatch + PARSEME



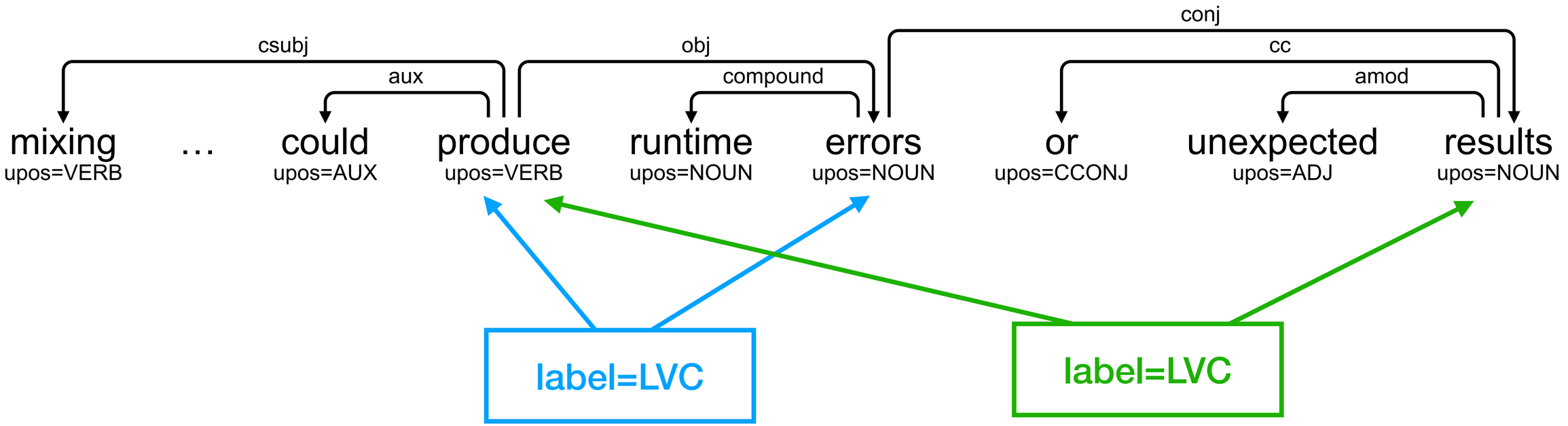
In **Grew**, every structure is a **graph**. We need to turn this structure into a **graph**

grewmatch + PARSEME

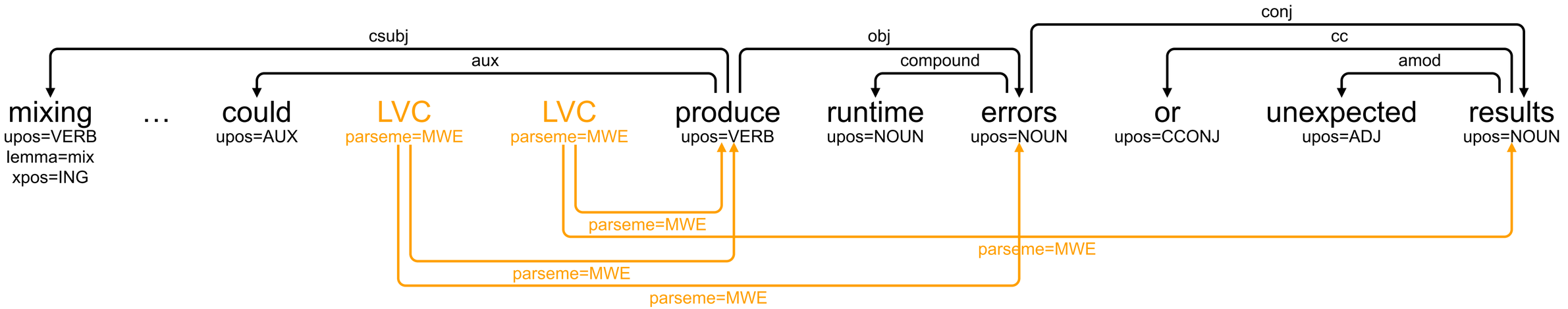
- Each **VMWE** is a new **node**
- **Edges** map each **VMWE** node to all **its tokens**



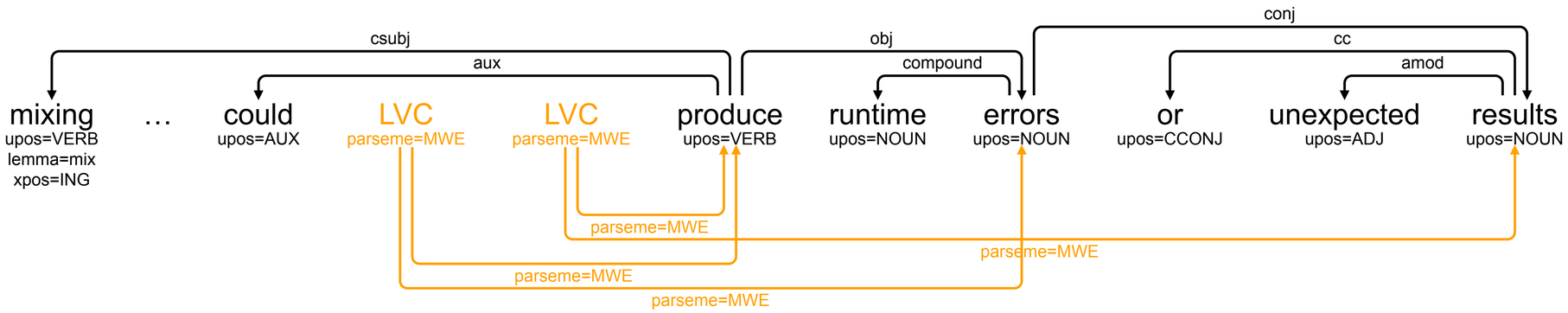
grewmatch + UD + PARSEME



grewmatch + UD + PARSEME



Grew-match request on it



- Request for a MWE node:

```
pattern { MWE [label="LVC"] }
```

Only MWE node have the **`label`** feature

- Request for a MWE edge:

```
pattern { MWE -[parseme=MWE]-> X}
```

```
pattern { MWE [label="LVC"]; MWE -> X}
```

An edge starting from an MWE node is necessarily an MWE edge

Explore how a treebank is annotated

```
pattern { MWE [label="VID"] }
```

Examples of VID usage

187 occurrences

```
pattern { MWE [label="VID"]; MWE -> V; V[upos=VERB] }
```

key clustering: `V.lemma`

lemmas of VERB in VID

197 occurrences

187 occurrences **197 occurrences**

Why do we observe this difference?

No verb ➔ 8
One verb ➔ 162
Two verb ➔ 16
Three verbs ➔ 1

197 = 1*3 + 16*2 + 162*1 + 8*0

Explore how a treebank is annotated

```
pattern { MWE1 [label] }
```

Find any MWE

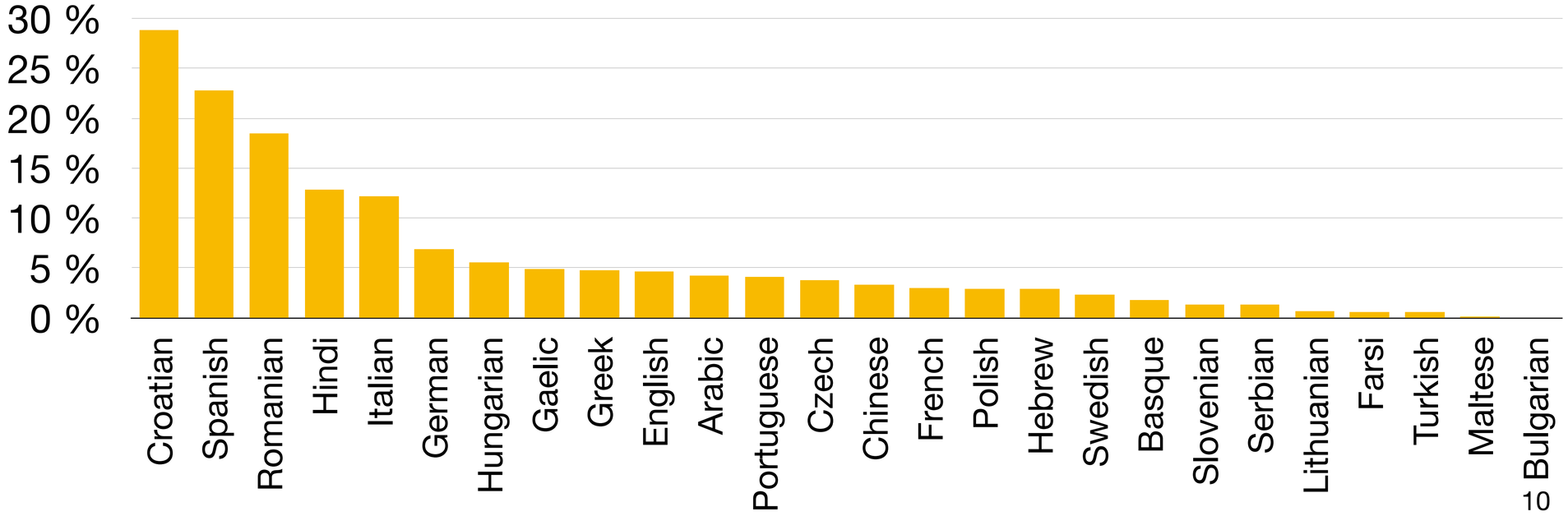
whether clustering:

```
MWE2 [label]; MWE1 -> X; MWE2 -> X
```

Check if there is an overlap with another MWE

4.67 % of VMWE overlap

in other languages?



Make linguistic observations

MVC usage in English

```
pattern { MWE [label="MVC"] }
```

Examples of MVC usage

```
pattern { MWE [label="MVC"] }
```

key clustering: `MWE.__out__`

number of edges starting from the node MWE

Sizes of MVC

```
pattern { MWE [label="MVC"]; MWE -> X1; MWE -> X2; X1 << X2 }
```

key clustering 1: `X1.lemma`

key clustering 2: `X2.lemma`

lemmas used in MVC

Make linguistic observations

is *make* + `obj` a VMWE?

yes or no?

```
pattern { N1 [lemma="make"]; N1 -[obj]-> N2 }
```

```
MWE [label]; MWE -> N1; MWE -> N2
```

60% No, 40% Yes

If yes, what VMWE `label`?

```
pattern {
  N1 [lemma="make"]; N1 -[obj]-> N2;
  MWE [label]; MWE -> N1; MWE -> N2
}
```

`MWE.label`

64 LVC.full | 8 VID | 3 LVC.cause

If yes, `lemma` of `obj`

```
pattern {
  N1 [lemma="make"]; N1 -[obj]-> N2;
  MWE [label]; MWE -> N1; MWE -> N2
}
```

`N2.lemma`

42 clusters

Corpus pre-annotation

13

Grew rules

- In **ArboratorGrew**, we can use **Grew** rules (like in UD/SUD conversion)
- It can be used
  - for automatic annotation of regular patterns
  - for automatic revision when we decide to change an existing annotation

Ex: make DET heads of NP

```
rule NP_to_DP {
  pattern {
    X[upos=DET]; Y[upos=NOUN];
    e:Y -[det]-> X
  }
  commands {
    del_edge e;
    shift X ==> Y;
    add_edge X -[comp:det]-> Y
  }
}
```

Ex: change tokenisation (don't ➔ do + n't)

```
rule r {
  pattern {
    X[form="don't"];
 }
  commands {
    add_node Y :> X;
    X.form = "do"; Y.form = "n't";
    Y.upos = PART;
 }
}
```

Using the parser for pre-annotation

- In **ArboratorGrew**, you can access to a parser (`BertForDeprel`)
  - Training on a part of already annotated data
  - Training on another treebanks of the same language, or a related language
  - Parsing with existing models