Syntax-based identification of light-verb constructions

Silvio Ricardo CORDEIRO
Marie CANDITO

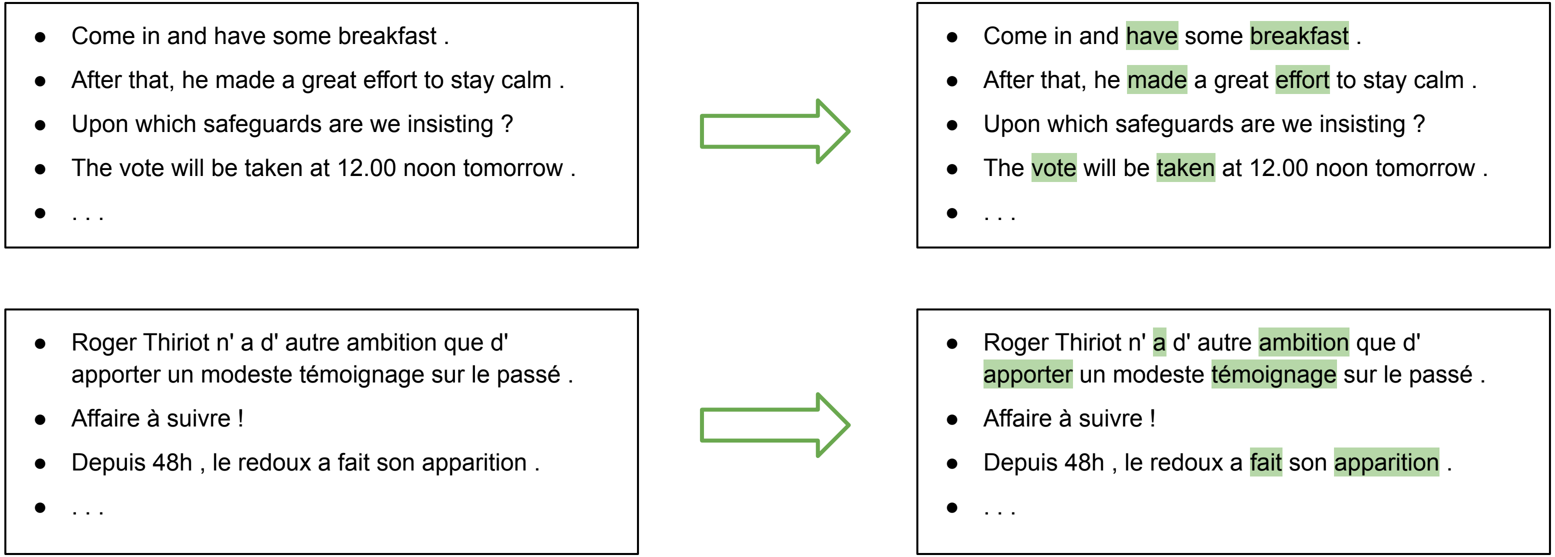
Overview

- Introduction
- Candidate identification
- Baseline and results
- Feature-based classifier and results
- Conclusions

Introduction

Introduction

- LVC identification in running text.



- Come in and have some breakfast .
- After that, he made a great effort to stay calm .
- Upon which safeguards are we insisting ?
- The vote will be taken at 12.00 noon tomorrow .
- . . .

- Come in and have some breakfast .
- After that, he made a great effort to stay calm .
- Upon which safeguards are we insisting ?
- The vote will be taken at 12.00 noon tomorrow .
- . . .

- Roger Thiriot n' a d' autre ambition que d' apporter un modeste témoignage sur le passé .
- Affaire à suivre !
- Depuis 48h , le redoux a fait son apparition .
- . . .

- Roger Thiriot n' a d' autre ambition que d' apporter un modeste témoignage sur le passé .
- Affaire à suivre !
- Depuis 48h , le redoux a fait son apparition .
- . . .

LVC properties

- LVC = Noun + verb that share a semantic argument;
- Verb is light;
  - Verb can be removed without an impact in semantics;
- Noun is predicative (event or state);
  - Similar nouns may form an LVC with the same verb;
- Syntactically flexible: gaps, overlap...

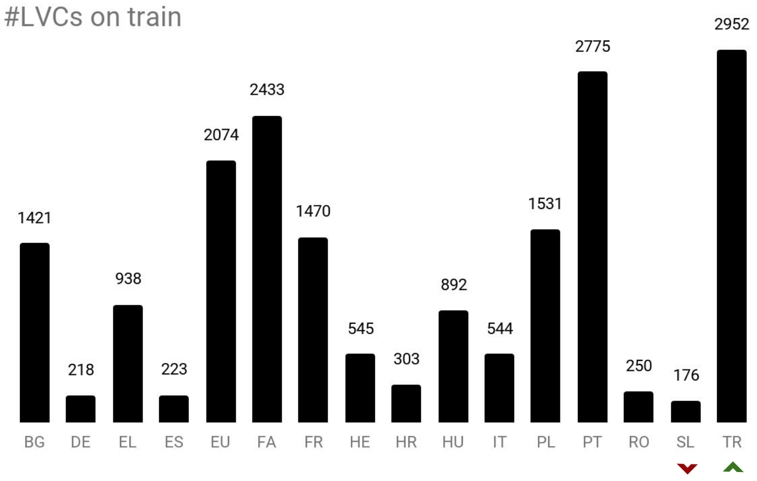
Example LVCs

- apporter témoignage
- faire apparition
- subir violence
- subir souffrance
- subir hémorragie

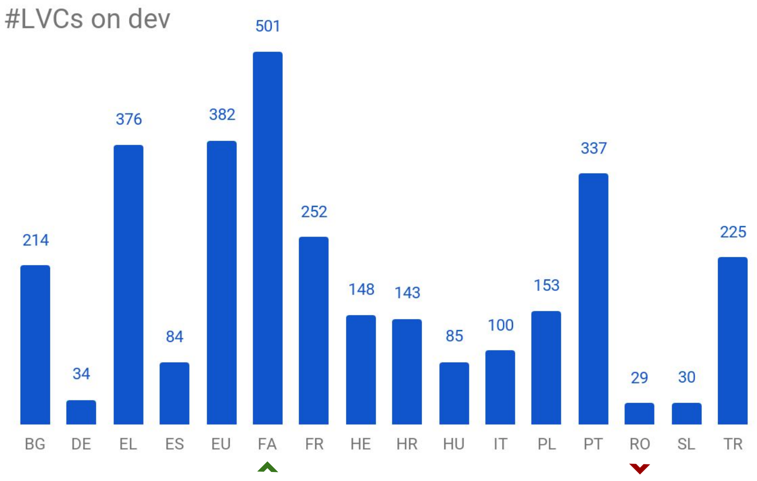
...

LVC statistics in PARSEME corpora

- PARSEME corpus with variable amounts of LVC annotations:

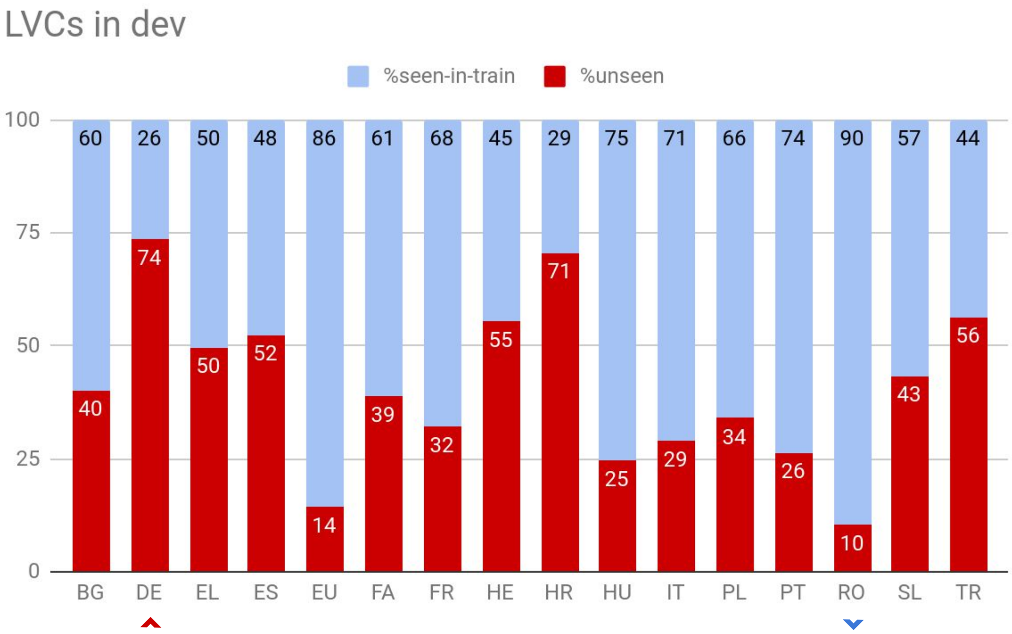


Average: 1171.6 LVCs



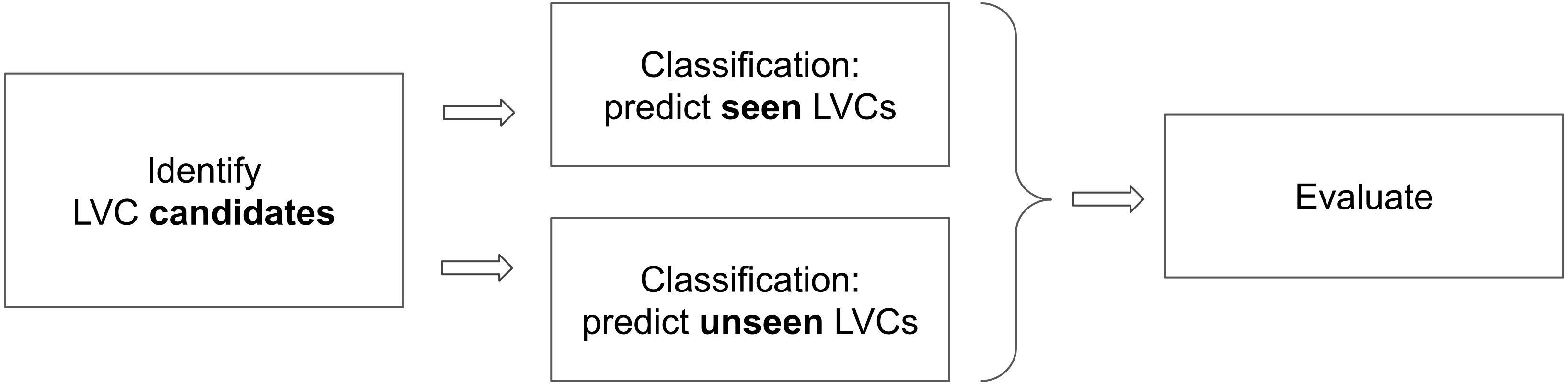
Average: 193.3 LVCs

LVC statistics in PARSEME corpora



Average: 40.7% unseen
MicroAvg: **38.5% unseen**

Proposed pipeline



Identify
LVC **candidates**

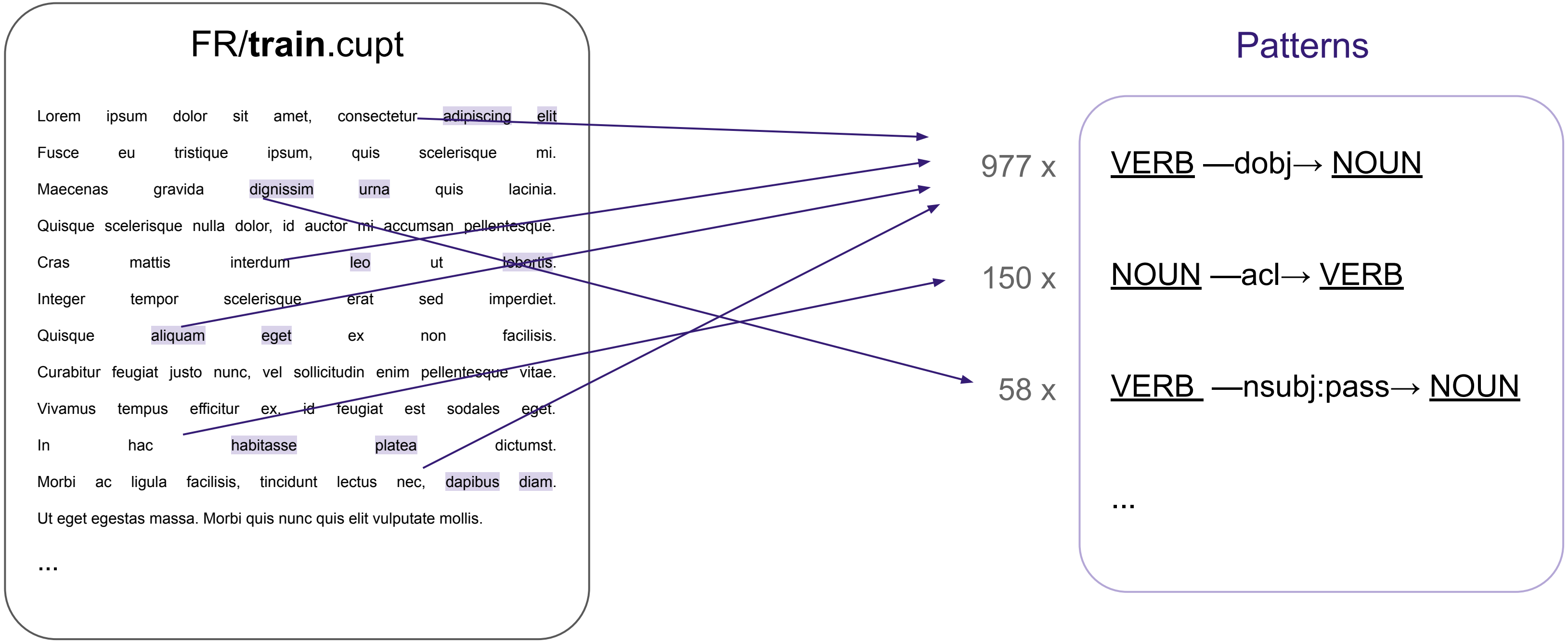
Classification:
predict **seen** LVCs

Classification:
predict **unseen** LVCs

Evaluate

LVC candidate identification

Pattern extraction



FR/**train**.cupt

Lorem ipsum dolor sit amet, consectetur adipiscing elit
Fusce eu tristique ipsum, quis scelerisque mi.
Maecenas gravida dignissim urna quis lacinia.
Quisque scelerisque nulla dolor, id auctor mi accumsan pellentesque.
Cras mattis interdum leo ut lobortis.
Integer tempor scelerisque erat sed imperdiet.
Quisque aliquam eget ex non facilisis.
Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.
Vivamus tempus efficitur ex, id feugiat est sodales eget.
In hac habitasse platea dictumst.
Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.
Ut eget egestas massa. Morbi quis nunc quis elit vulputate mollis.
...

Patterns

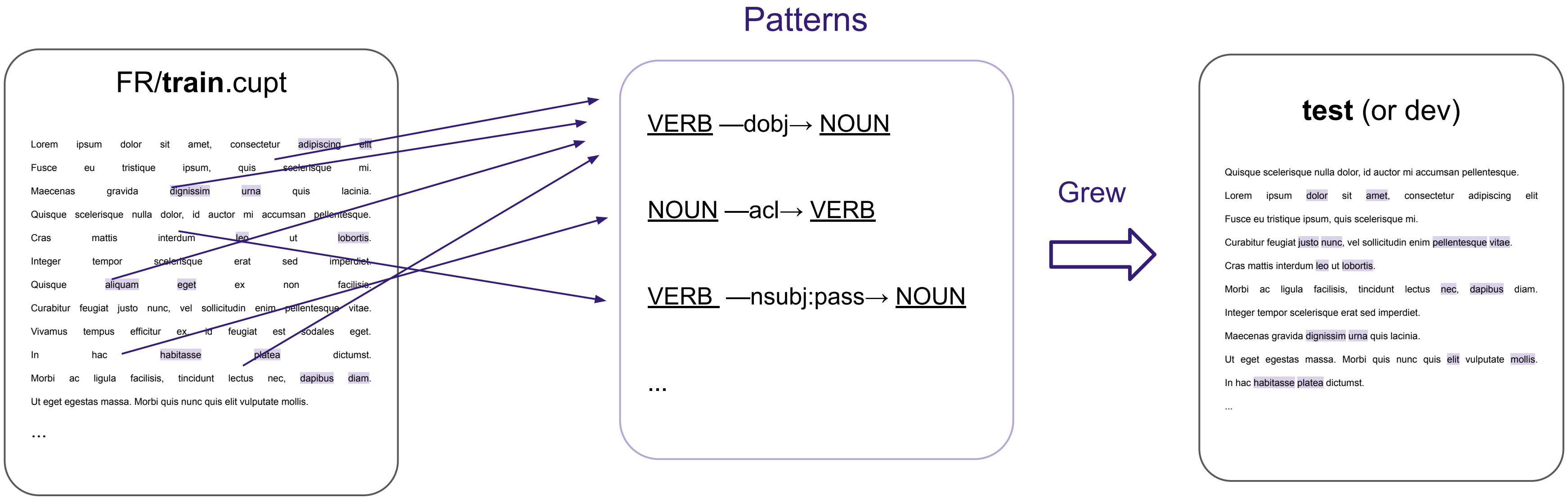
977 x VERB —dobj→ NOUN

150 x NOUN —acl→ VERB

58 x VERB —nsubj:pass→ NOUN

...

Candidate identification



FR/**train**.cupt

Lorem ipsum dolor sit amet, consectetur adipiscing elit
Fusce eu tristique ipsum, quis scelerisque mi.
Maecenas gravida dignissim urna quis lacinia.
Quisque scelerisque nulla dolor, id auctor mi accumsan pellentesque.
Cras mattis interdum leo ut lobortis.
Integer tempor scelerisque erat sed imperdiet.
Quisque aliquam eget ex non facilisis.
Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.
Vivamus tempus efficitur ex id feugiat est sodales eget.
In hac habitasse platea dictumst.
Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.
Ut eget egestas massa. Morbi quis nunc quis elit vulputate mollis.
...

Patterns

VERB —dobj→ NOUN

NOUN —acl→ VERB

VERB —nsubj:pass→ NOUN

...

test (or dev)

Quisque scelerisque nulla dolor, id auctor mi accumsan pellentesque.
Lorem ipsum dolor sit amet, consectetur adipiscing elit
Fusce eu tristique ipsum, quis scelerisque mi.
Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.
Cras mattis interdum leo ut lobortis.
Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.
Integer tempor scelerisque erat sed imperdiet.
Maecenas gravida dignissim urna quis lacinia.
Ut eget egestas massa. Morbi quis nunc quis elit vulputate mollis.
In hac habitasse platea dictumst.
...

]

LVC candidate coverage

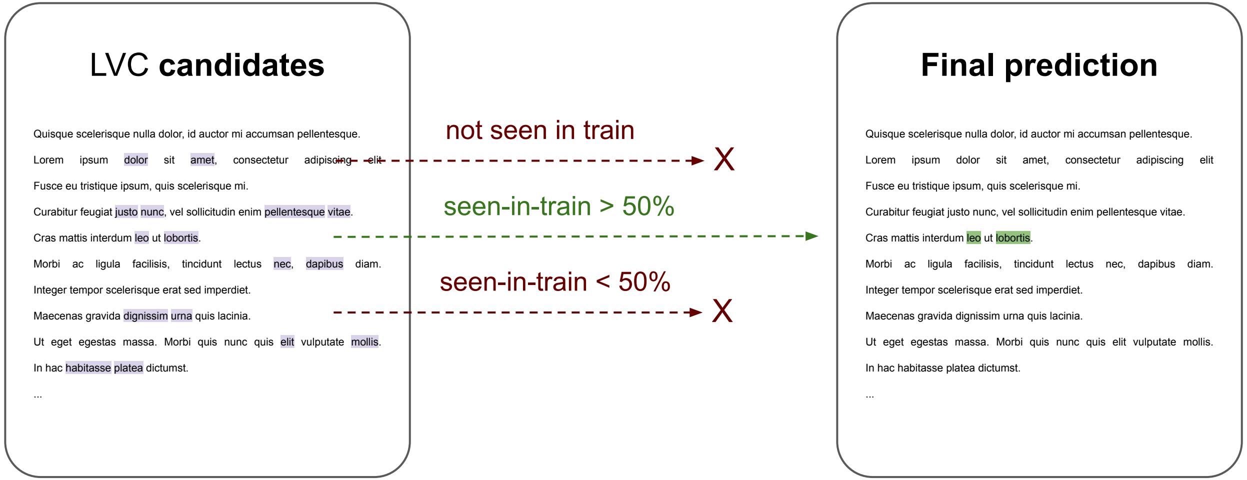
- Taking all patterns stemming from *at least 2 occurrences*

Coverage	BG	DE	EL	ES	EU	FA	FR	HE	HR	HU	IT	PL	PT	RO	SL	TR	Avg	MicroAvg
On seen	98	**100**	**100**	95	94	91	99	**82**	93	86	90	97	95	96	**100**	98	94.6	**94.6**
On unseen	73	84	91	**98**	**98**	90	91	78	96	86	**72**	90	93	**33**	92	95	85.2	**89.5**

Seen+Unseen	88	88	96	96	95	90	96	**80**	95	86	85	95	94	90	**97**	96	91.7	**92.6**

Baseline

Baseline for LVCs seen in train



LVC **candidates**

Quisque scelerisque nulla dolor, id auctor mi accumsan pellentesque.
Lorem ipsum dolor sit amet, consectetur adipiscing elit
Fusce eu tristique ipsum, quis scelerisque mi.
Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.
Cras mattis interdum leo ut lobortis.
Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.
Integer tempor scelerisque erat sed imperdiet.
Maecenas gravida dignissim urna quis lacinia.
Ut eget egestas massa. Morbi quis nunc quis elit vulputate mollis.
In hac habitasse platea dictumst.
...

not seen in train

X

seen-in-train > 50%

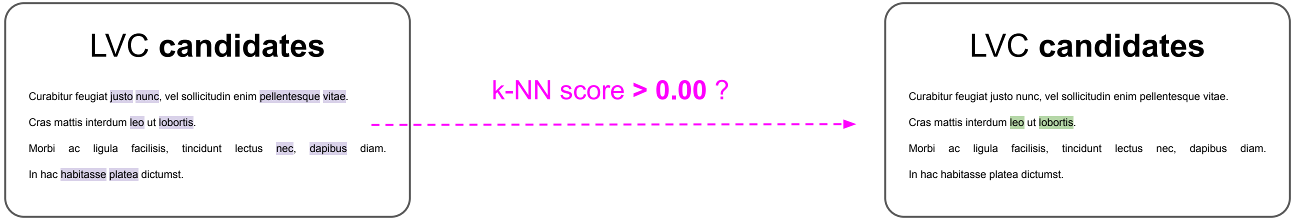
seen-in-train < 50%

X

Final prediction

Quisque scelerisque nulla dolor, id auctor mi accumsan pellentesque.
Lorem ipsum dolor sit amet, consectetur adipiscing elit
Fusce eu tristique ipsum, quis scelerisque mi.
Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.
Cras mattis interdum leo ut lobortis.
Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.
Integer tempor scelerisque erat sed imperdiet.
Maecenas gravida dignissim urna quis lacinia.
Ut eget egestas massa. Morbi quis nunc quis elit vulputate mollis.
In hac habitasse platea dictumst.
...

Baseline for unseen LVCs



LVC **candidates**

Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.

Cras mattis interdum leo ut lobortis.

Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.

In hac habitasse platea dictumst.

k-NN score **> 0.00** ?

LVC **candidates**

Curabitur feugiat justo nunc, vel sollicitudin enim pellentesque vitae.

Cras mattis interdum leo ut lobortis.

Morbi ac ligula facilisis, tincidunt lectus nec, dapibus diam.

In hac habitasse platea dictumst.

k-Nearest Neighbors of *enseignant*
in LVC candidate **mener enseignant**

noun	**cosine**	**Annotated as LVC with *mener*?**
inspecteur	0.47	NO
patient	0.37	NO
institut	0.35	NO
conduite	0.29	YES
gouvernement	0.27	NO
	...	

4-NN score
= Σ {−0.47, −0.37, −0.35, +0.29}
= **−0.90**

Results for the baseline

- Evaluated on the *test* datasets:

Baseline F1	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**EN**	**HI**	**LT**	**Avg**	**MicroAvg**
Seen (>50%)	74	67	84	61	88	84	87	70	86	**92**	86	86	90	83	72	53	64	90	**48**	77.1	**80.1**
Unseen (kNN)	14	9	23	17	16	21	34	5	18	24	13	19	28	**0**	5	36	23	**44**	8	17.5	**22.5**
Seen+Unseen	53	**26**	62	36	**81**	64	62	30	45	77	63	60	74	69	34	44	31	68	28	55.0	**57.3**

Baseline vs shared-task systems

- Evaluated on the *test* datasets:

F1	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**EN**	**HI**	**LT**	**Avg**	**MicroAvg**
Baseline	53	**26**	62	36	**81**	64	62	30	45	77	63	60	74	69	34	44	31	68	28	55.0	**57.3**
SHOMA	50	**0**	60	22	79	78	51	43	24	59	46	51	70	**86**	28	64	2	72	29	50.8	**56.4**
TRAVERSAL	44	**15**	47	26	70	65	52	30	32	68	51	52	62	**73**	38	44	18	62	23	48.3	**50.2**

F1 (unseen)	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**EN**	**HI**	**LT**	**Avg**	**MicroAvg**
Baseline	14	9	23	17	16	21	34	5	18	24	13	19	28	**0**	5	36	23	**44**	8	17.5	**22.5**
SHOMA	21	**0**	36	13	35	**62**	37	19	19	14	4	22	35	29	**0**	50	3	53	8	24.8	**30.6**
TRAVERSAL	8	**0**	18	10	11	41	31	5	21	23	**0**	20	24	**0**	**0**	23	14	**42**	1	14.7	**20.4**

Feature-based classifier

Features for each LVC candidate

- F_1: Grew pattern (one-hot)
- F_2: Seen-in-train fraction
- F_3: POS tag of verb and noun (one-hot)
- F_4: Length of the LVC (one-hot)

- F_E: Word embeddings (fasttext, 300 dims)
- F_N: k-NN score

- F_C: Binary contextual features from UD columns
  - e.g. "verb has a dependent with VerbForm=Inf"
  - e.g. "noun has a dependent with DEPREL=nmod"
  - We take the top *t* features with highest mutual information

Classifiers

- Support-vector machine (SVM):
    - RBF kernel;
    - 3-fold gridsearch;
        - C $\in$ {1, 10, 20, 50, 100;
        - Gamma $\in$ {0.5, 0.1, 0.05, 0.01}.

- Feed-forward neural network (FFN):
    - 100-neuron hidden layer;
    - Optimizer: SGD (LR=0.01);
    - Loss: negative log-likelihood;
    - Activation function: tanh;
    - Dropout: 50%;
    - Minibatches of size 4.

Results

- Evaluated on the *test* datasets:

F1	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**EN**	**HI**	**LT**	**Avg**	**MicroAvg**
Baseline	53	**26**	62	36	**81**	64	62	30	45	77	63	60	74	69	34	44	31	68	28	**55.0**	**57.3**
SVM	61	40	66	35	79	77	65	41	44	**81**	70	71	78	67	63	61	**26**	77	28	**62.3**	**63.3**
FFN	53	26	43	36	74	74	51	**21**	42	75	44	60	68	57	26	56	40	**78**	30	**50.5**	**56.3**
SHOMA	50	**0**	60	22	79	78	51	43	24	59	46	51	70	**86**	28	64	2	72	29	**50.8**	**56.4**
TRAVERSAL	44	**15**	47	26	70	65	52	30	32	68	51	52	62	**73**	38	44	18	62	23	**48.3**	**50.2**

Results

- Evaluated on the *test* datasets:

F1 (unseen)	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**EN**	**HI**	**LT**	**Avg**	**MicroAvg**
Baseline	14	9	23	17	16	21	34	5	18	24	13	19	28	**0**	5	36	23	**44**	8	**17.5**	**22.5**
SVM	17	24	34	**7**	17	57	29	**7**	17	36	6	39	39	**67**	13	43	19	61	**7**	**28.3**	**31.0**
FFN	20	18	18	20	19	45	25	6	16	29	12	33	31	**0**	7	34	33	**64**	15	**20.9**	**29.3**
SHOMA	21	**0**	36	13	35	**62**	37	19	19	14	4	22	35	29	**0**	50	3	53	8	**24.8**	**30.6**
TRAVERSAL	8	**0**	18	10	11	**41**	31	5	21	23	**0**	20	24	**0**	**0**	23	14	42	1	**14.7**	**20.4**

Conclusions

Conclusions

- LVC identification on PARSEME 1.1 data;
- Seen vs unseen: different subtasks;
- Strong baseline, beats the best systems in the shared-task;
- SVM results surpass the best system by 7 percentage points;
- Results for *unseen* LVCs still much lower than results for *seen*.

Syntax-based identification of light-verb constructions

Thank you

Silvio Ricardo CORDEIRO
Marie CANDITO

Additional slides

LVC statistics in PARSEME corpora

- PARSEME corpus with variable amounts of LVC annotations:

#LVCs	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**Avg**
On train	1421	218	938	223	2074	2433	1470	545	303	892	544	1531	2775	250	**176**	**2952**	**1171.6**
On dev	214	34	376	84	382	**501**	252	148	143	85	100	153	337	**29**	30	225	**193.3**

- Seen vs unseen LVCs:

%seen	**BG**	**DE**	**EL**	**ES**	**EU**	**FA**	**FR**	**HE**	**HR**	**HU**	**IT**	**PL**	**PT**	**RO**	**SL**	**TR**	**Avg**	**MicroAvg**
On dev	60	**26**	50	48	86	61	68	45	29	75	71	66	74	**90**	57	44	59.3	**61.5**