Do we Name the Languages we Study?

The #BenderRule in LREC and ACL articles

Fanny Ducel[†], Karën Fort[$⋆], Gaël Lejeune[†], Yves Lepage[‡]

[†]STIH, Sorbonne Université, France [$]Sorbonne Université, France
[⋆] Université de Lorraine, CNRS, Inria, LORIA, France
[‡]Waseda University, Japan

ducelfanny@gmail.com, karen.fort@loria.fr,
gael.lejeune@sorbonne-universite.fr, yves.lepage@waseda.jp

Introduction

Context: NLP4NLP

- *Which languages*, and how many, are studied in NLP articles?
- Languages studied should at the very least be **stated**.
- This has linguistic, sociological and ethical issues

Context: NLP4NLP

- *Which languages*, and how many, are studied in NLP articles?
- Languages studied should at the very least be **stated**.
- This has linguistic, sociological and ethical issues

Is it obvious ?

> *"Do state the name of the language that is being studied, even if it's English. Acknowledging that we are working on a particular language foregrounds the possibility that the techniques may in fact be language specific. Conversely, neglecting to state that the particular data used were in, say,* ***English, gives [a] false veneer of language-independence to the work****." [Bender, 2011]*

Context: NLP4NLP

- *Which languages*, and how many, are studied in NLP articles?
- Languages studied should at the very least be **stated**.
- This has linguistic, sociological and ethical issues

Is it obvious ?

> *"Do state the name of the language that is being studied, even if it's English. Acknowledging that we are working on a particular language foregrounds the possibility that the techniques may in fact be language specific. Conversely, neglecting to state that the particular data used were in, say,* ***English, gives [a] false veneer of language-independence to the work****." [Bender, 2011]*
>
> ***"Always name the language(s) you're working on." [Bender, 2019]***

Contributions

- Empirically examining how much the #BenderRule is applied.

Contributions

- Empirically examining how much the #BenderRule is applied.

Experimental setup

- A corpus of 14,000 articles (LREC and ACL)
- A subcorpus of 1,000 articles manually annotated:
  - is the #BenderRule applicable
  - if yes, is it applied or not

Contributions

- Empirically examining how much the #BenderRule is applied.

Experimental setup

- A corpus of 14,000 articles (LREC and ACL)
- A subcorpus of 1,000 articles manually annotated:
  - is the #BenderRule applicable
  - if yes, is it applied or not
- we trained a classifier from this data
- we applied this classifier to the remainder of the data
- we analyzed the results from a contrastive and diachronic perspective

Methodology

Corpus building

A corpus of articles in English: LREC (2000-2020) & ACL (1979-2020)

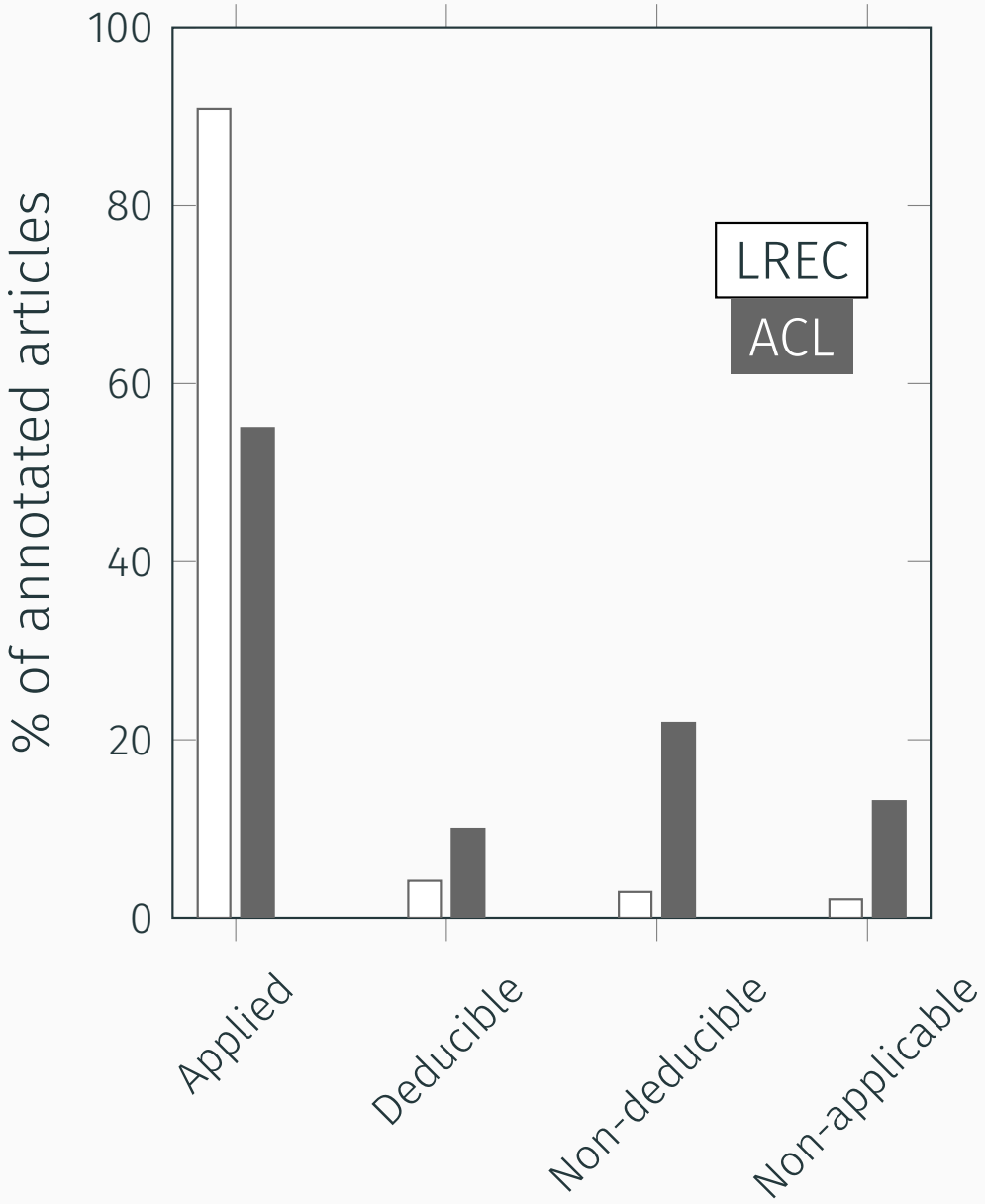
	LREC	ACL
Text files	0	4,867
Converted files	6,715	2,419
Excluded files	46	24
Total	6,669	7,262

Table 1: Composition of the corpus.

Manual Annotation of 970 articles

	#BenderRule applicable	#BenderRule applied	Language mentioned	Resource mentioned
Non-applic.	-	-	(*)	(*)
Non-ded.	+	-	-	-
Deducible	+	-	-	+
Applied	+	+	+	(*)

Results of Manual Annotation (%)



Classifiers

Baseline Classifier using Pattern-Matching, Class_{PM}

- #BenderRule applied: 1+ language name found in the text

Classification using Machine Learning, Class_{ML}

- Trained on sentences from manually annotated ACL articles
- #BenderRule applied: 1+ sentence classified as such

Results

Performance of Classifiers

	$Class_{PM}$					
	LREC (correlation = 0.634)			ACL (correlation = 0.509)		
	Prec.	F-meas.	# instances	Prec.	F-meas.	# instances
Applied	0.894	0.941	440	0.729	0.835	231
Deducible	**1.000**	**0.788**	20	**1.000**	**0.746**	42
N/A	**0.792**	0.551	90	0.741	0.543	147
Macro avg.	**0.895**	0.760	550	0.823	**0.708**	420
	$Class_{ML}$					
	LREC (correlation = **0.671**)			ACL (correlation = **0.741**)		
	Prec.	F-meas.	# instances	Prec.	F-meas.	# instances
Applied	**0.908**	**0.946**	440	**0.856**	**0.911**	231
Deducible	**1.000**	0.750	20	**1.000**	0.444	42
N/A	0.767	**0.613**	90	**0.752**	**0.747**	147
Macro avg.	0.892	**0.770**	550	**0.869**	0.701	420

Table 2: Performance of the classifiers and Spearman correlation with Annotator 0 on the two sub-corpora

.

Classification Results: Machine Learning (in %)

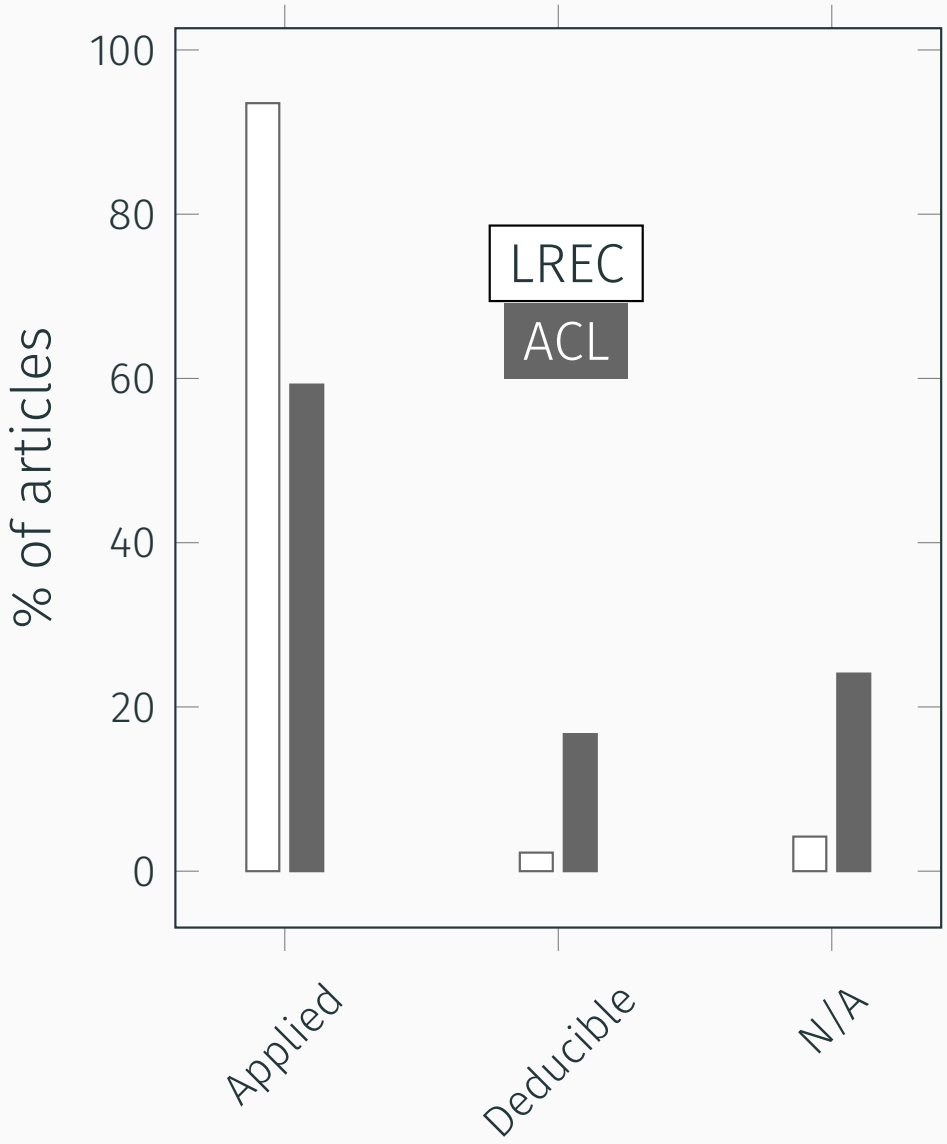


Figure 1: Results, in %, of automatic classification by Class$_{ML}$

Diachronic Study: The #BenderRule over the Years

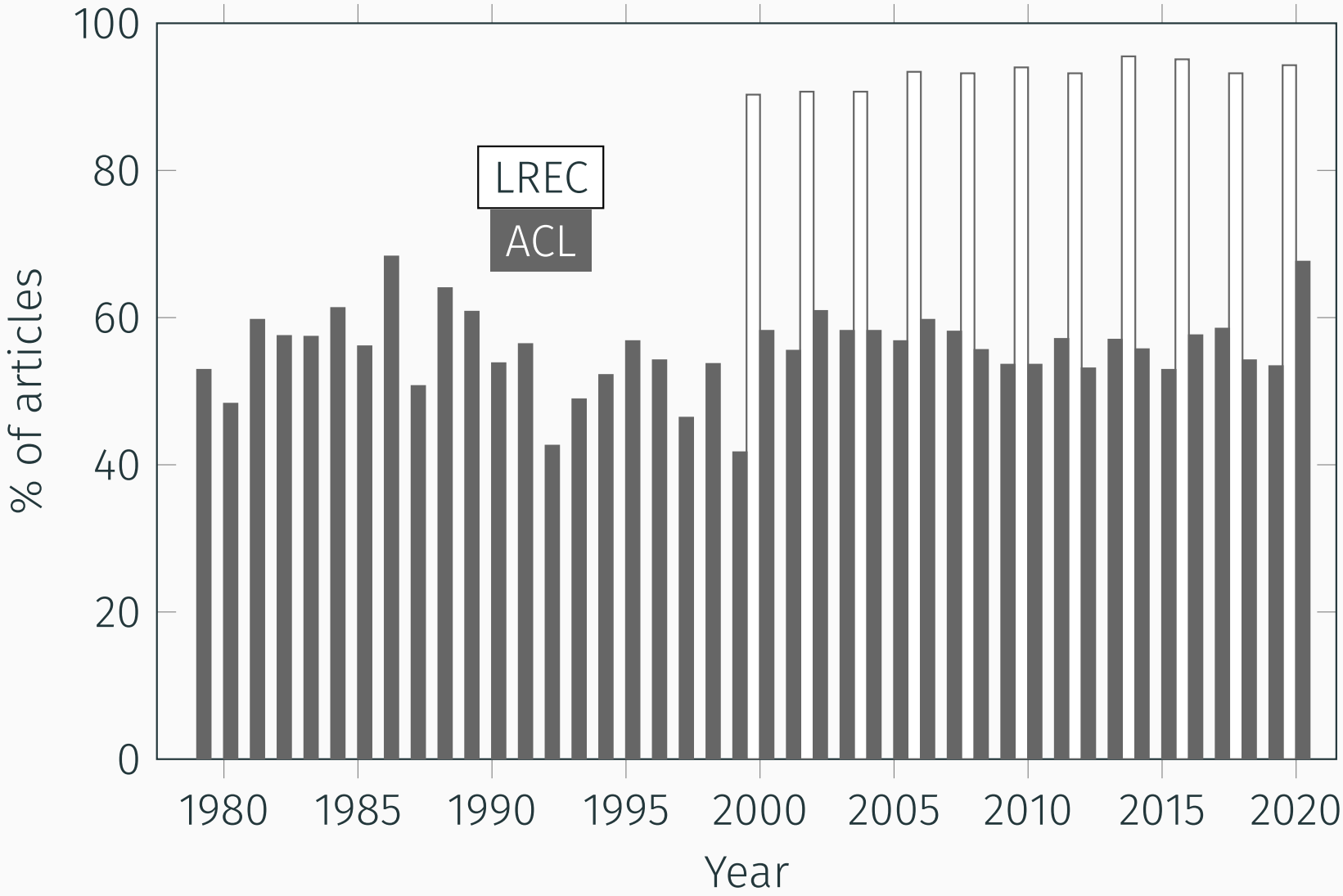


Figure 2: Number of articles applying the #BenderRule, in % in each edition of both conferences

Contrastive Study: How many Languages Mentioned?

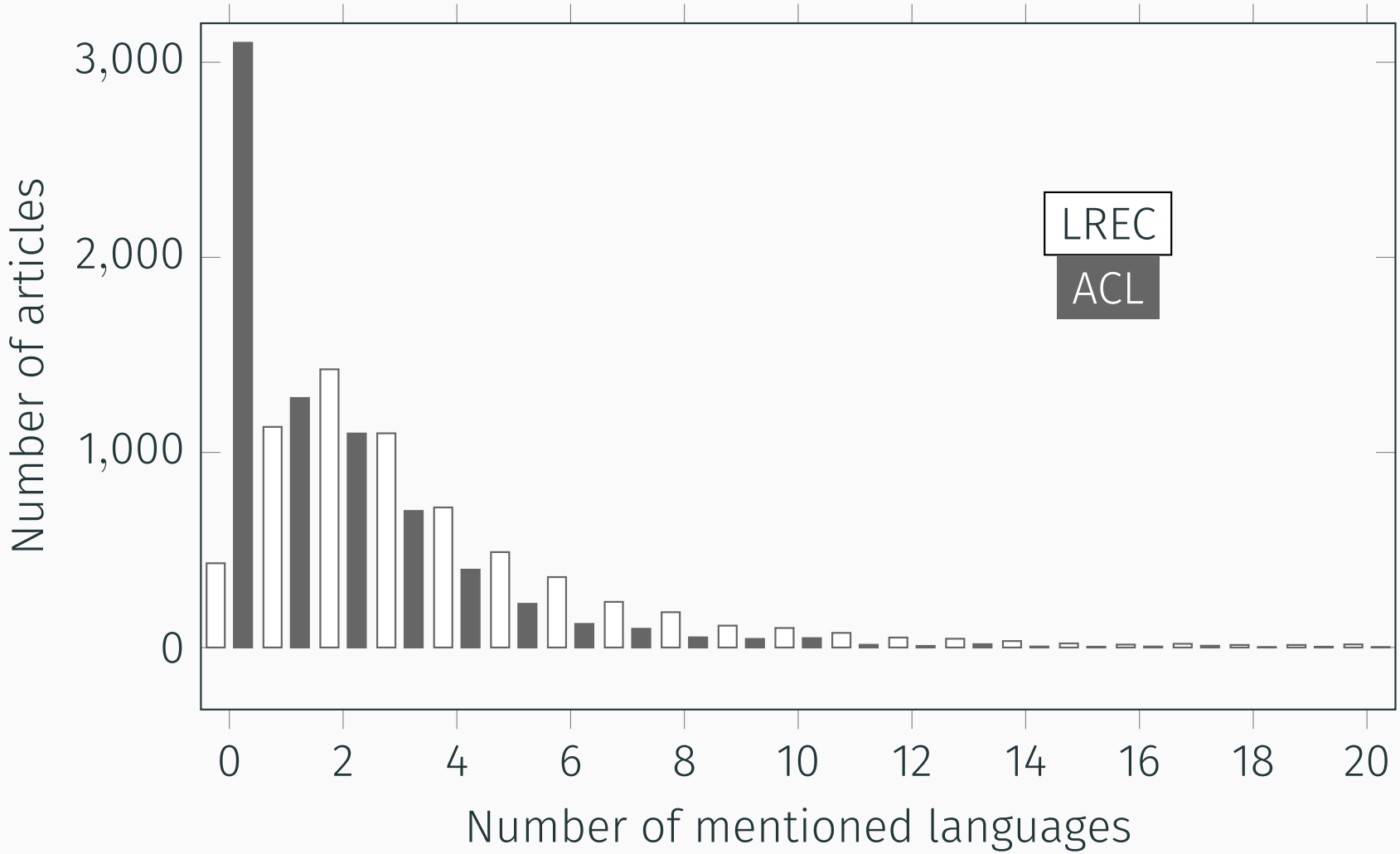


Figure 3: Distribution of number of articles per number of languages mentioned

Contrastive Study: Languages for which the #BenderRule is the most Applied

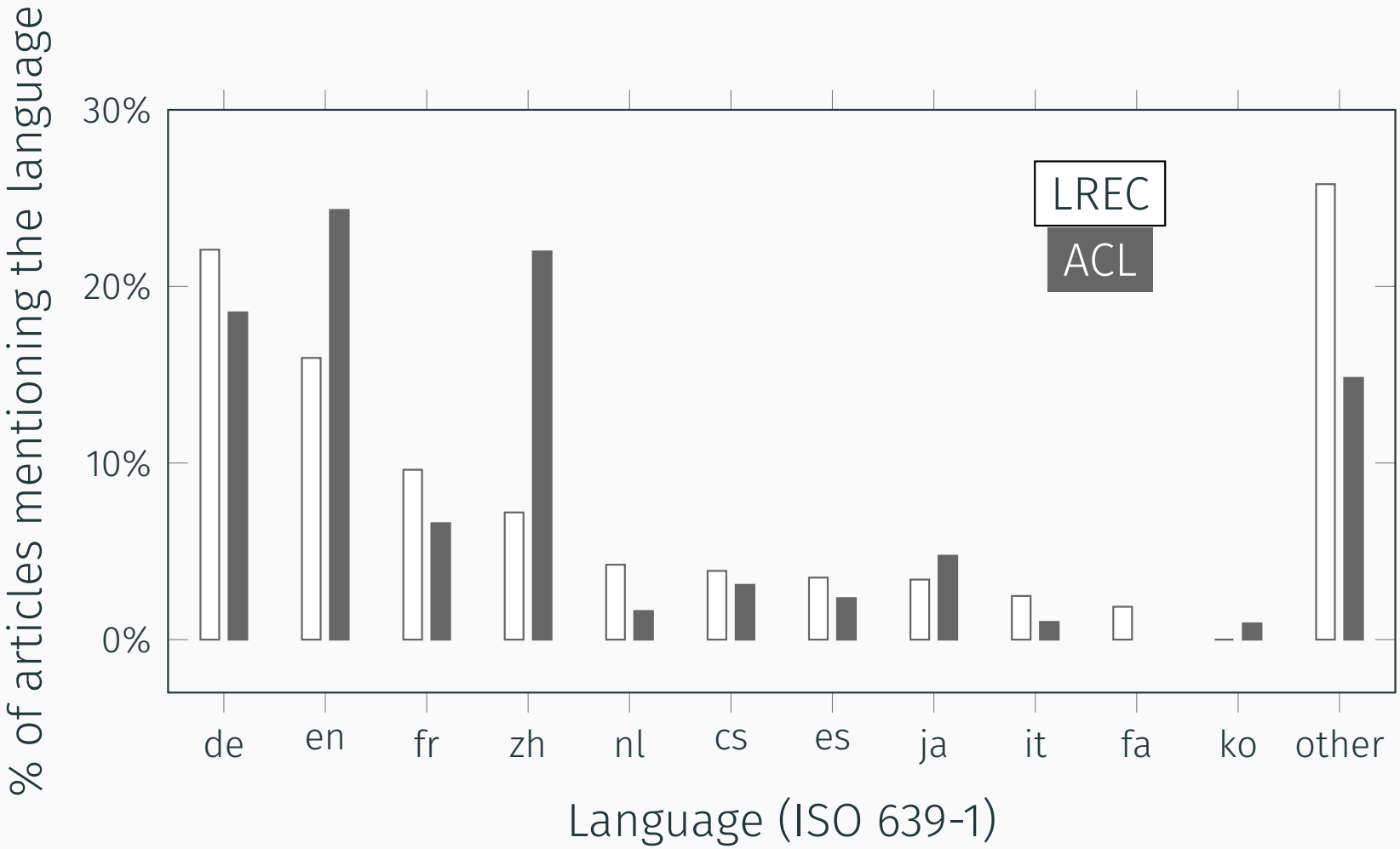


Figure 4: % of articles mentioning a given language in each corpus, over all articles applying the #BenderRule

Potential Limitations of the Study

Potential Limitations of the Study

- Questionable annotations
- Non-exhaustive list of languages
- LREMap as a unique reference for deducing the language studied
- No measure of influence of Bender's blog introducing #BenderRule (2019)
  $\rightarrow$ rerun these experiments in a few years

Conclusion

Conclusion

- **Diachronic study**:
  - No significant change over time
- **Contrastive study**
  - LREC: + mentions, linguistic diversity, multilingual works than ACL
- In both corpora: #BenderRule not applied $\Rightarrow$ **English**
- Code and resources used freely available on GitHub
- Replicate the experiment on articles from other NLP conferences

References


Bender, E. (2019).
The #BenderRule: On naming the languages we study and why it matters.
The Gradient.


Bender, E. M. (2011).
On achieving and evaluating language-independence in NLP.
Linguistic Issues in Language Technology, 6(3).