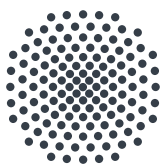
Lexicalized vs. Delexicalized Parsing in Low-Resource Scenarios

Agnieszka Falenska Özlem Çetinoğlu

Institute for Natural Language Processing



Universität Stuttgart

Table of Contents

Origins

- It all started with the CoNLL-ST. (Zeman et al., 2017)
- Particularly interesting: low-resource scenarios.

Low-resource scenarios of the CoNLL-ST

- Languages with very small training treebanks.

Surprise languages		
bxr	Buryat	19 sent.
kmr	Kurmanji	20 sent.
sme	North Sami	20 sent.
hsb	Upper Sorbian	23 sent.

Small languages		
kk	Kazakh	31 sent.
ug	Uyghur	100 sent.
ga	Irish	566 sent.
uk	Ukrainian	863 sent.
la	Latin	1334 sent.

Why are they interesting?

It is not obvious how to approach them.

- When you have a big treebank – train a parser.
- When there is no treebank – apply cross-lingual methods.

Source: English

Target: Polish

- What if you have a treebank but it is very small?

Why are they interesting?

It is not obvious how to approach them.

- When you have a big treebank – train a parser.
- When there is no treebank – apply cross-lingual methods.

Source: English

Target: Polish

- What if you have a treebank but it is very small?

Research questions

Given a language with a small treebank:

- monolingual or cross-lingual method?
- how to decide?
  - depending on the treebank size?
  - depending on the POS accuracy?
  - depending on the language itself?

Given a new language without a treebank:

- can we achieve useful results with cross-lingual methods ...
- ... or the only way to parse it is to annotate new trees?

Research questions

Given a language with a small treebank:

- monolingual or cross-lingual method?
- how to decide?
  - depending on the treebank size?
  - depending on the POS accuracy?
  - depending on the language itself?

Given a new language without a treebank:

- can we achieve useful results with cross-lingual methods ...
- ... or the only way to parse it is to annotate new trees?

Research questions

Given a language with a small treebank:

- monolingual or cross-lingual method?
- how to decide?
  - depending on the treebank size?
  - depending on the POS accuracy?
  - depending on the language itself?

Given a new language without a treebank:

- can we achieve useful results with cross-lingual methods ...
- ... or the only way to parse it is to annotate new trees?

Table of Contents

Low-resource scenarios of the CoNLL-ST

	Surprise languages	Small languages
Training treebank	very small	small
Parallel data	no	no/small
POS gold data:	external[1]	treebank
POS accuracy:	good	varies

[1] simulated for the CoNLL-ST

Low-resource scenarios of the CoNLL-ST

	Surprise languages	Small languages
Training treebank	very small	small
Parallel data	no	no/small
POS gold data:	external[1]	treebank
POS accuracy:	good	varies

[1] simulated for the CoNLL-ST

Generalized low-resource scenarios beyond CoNLL-ST

	ExternalPOS	TreebankPOS
Training treebank	very small to small	very small to small
Parallel data	no	no
POS gold data	external	treebank
POS accuracy:	good	varies
Test cases:		

Generalized low-resource scenarios beyond CoNLL-ST

	ExternalPOS	TreebankPOS
Training treebank	very small to small	very small to small
Parallel data	no	no
POS gold data	external	treebank
POS accuracy:	good	varies
Test cases:	surprise languages	small languages

Simulated low-resource scenarios

- Test languages:
  - surprise/small languages of CoNLL-ST

- Experiments, analysis:
  - simulate low-resource with 41 big CoNLL-ST treebanks
  - artificial training samples: starting with 100 tokens
    - for ExternalPOS– sample trees
    - for TreebankPOS– trees and POS tags

Simulated low-resource scenarios

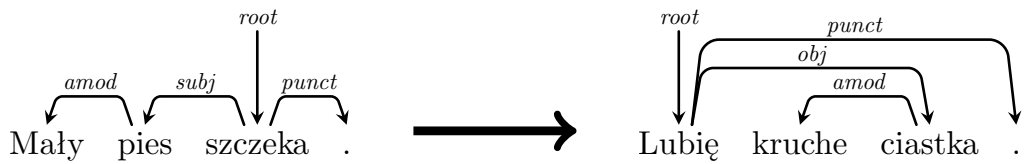
- Test languages:
  - surprise/small languages of CoNLL-ST

- Experiments, analysis:
  - simulate low-resource with 41 big CoNLL-ST treebanks
  - artificial training samples: starting with 100 tokens
    - for ExternalPOS– sample trees
    - for TreebankPOS– trees and POS tags

Monolingual method

Method (LEX):

- train a parser on the target treebank (even 10 sentences)



Source: Polish Target: Polish

Tool:

- transition-based parser (Björkelund and Nivre, 2015)
- default features and parameters

Cross-lingual method

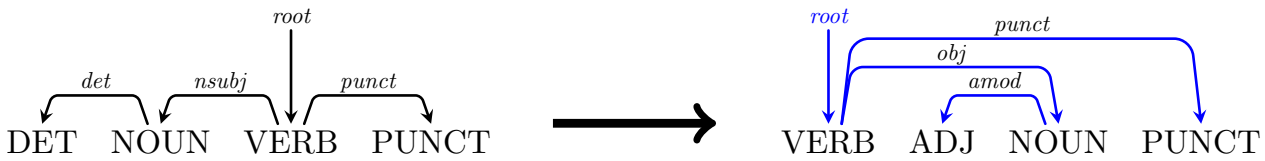
- Most of the cross-lingual techniques need parallel data.
- Alternative: delexicalized parsing (Zeman and Resnik, 2008)

Source: English

Target: Polish

Cross-lingual method

- Most of the cross-lingual techniques need parallel data.
- Alternative: delexicalized parsing (Zeman and Resnik, 2008)



Source: English

Target: Polish

Cross lingual method – DELEX

Method (DELEX):

- Delexicalized multisource model transfer and weighted blending.

  (Sagae and Lavie, 2006), (Rosa and Žabokrtský, 2015), (Agić, 2017)

- Details, colourful plots and animation - backup slides.

- Samples of target trees - similarities between sources and targets.

Cross lingual method – DELEX

Method (DELEX):

- Delexicalized multisource model transfer and weighted blending.
  (Sagae and Lavie, 2006), (Rosa and Žabokrtský, 2015), (Agić, 2017)
- Details, colourful plots and animation - backup slides.

- Samples of target trees - similarities between sources and targets.

Cross lingual method – DELEX

Method (DELEX):

- Delexicalized multisource model transfer and weighted blending.
  (Sagae and Lavie, 2006), (Rosa and Žabokrtský, 2015), (Agić, 2017)
- Details, colourful plots and animation - backup slides.

- Samples of target trees - similarities between sources and targets.

Methodology – few more details

- Universal Dependencies v2.0 treebanks (Nivre et al., 2016)
- Two settings: ExternalPOS and TreebankPOS
- 46 source languages
- 41 simulated targets

- Goal: compare LEX and DELEX methods

Methodology – few more details

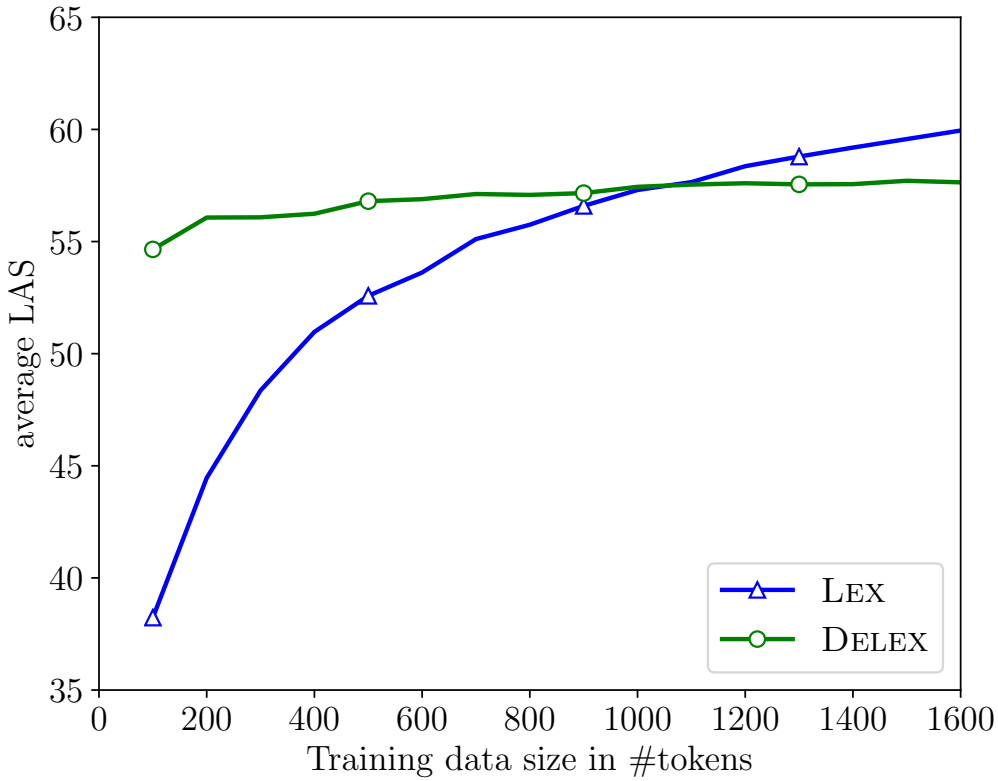
- Universal Dependencies v2.0 treebanks (Nivre et al., 2016)
- Two settings: ExternalPOS and TreebankPOS
- 46 source languages
- 41 simulated targets

- Goal: compare LEX and DELEX methods

Table of Contents

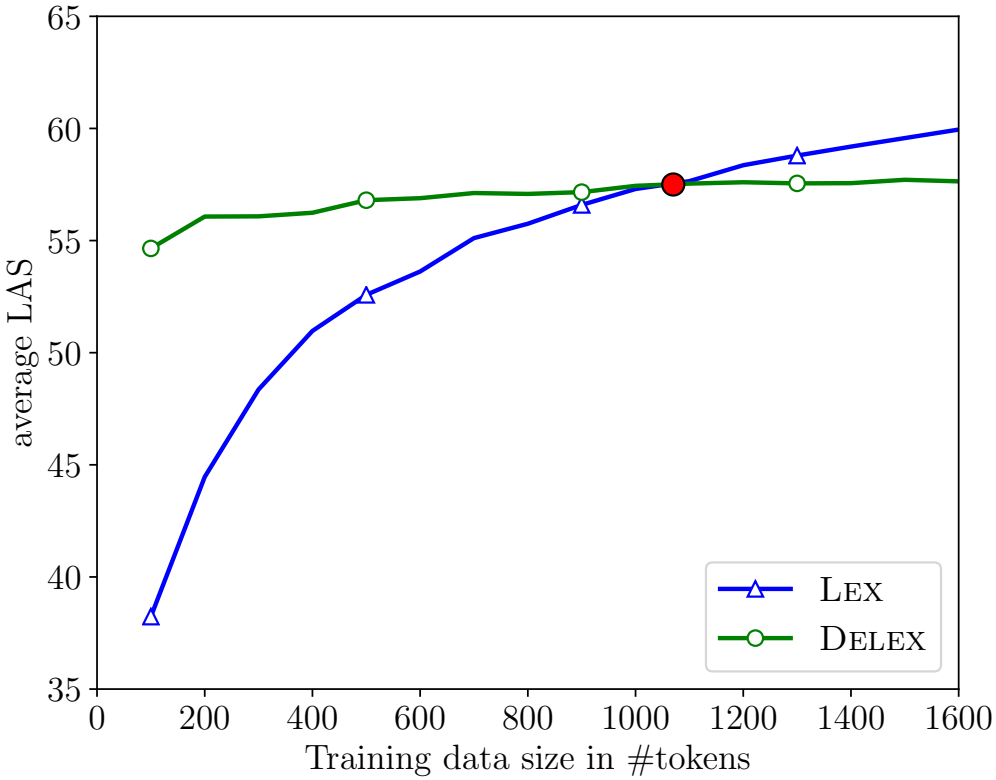
Table of Contents

Average results across 41 targets



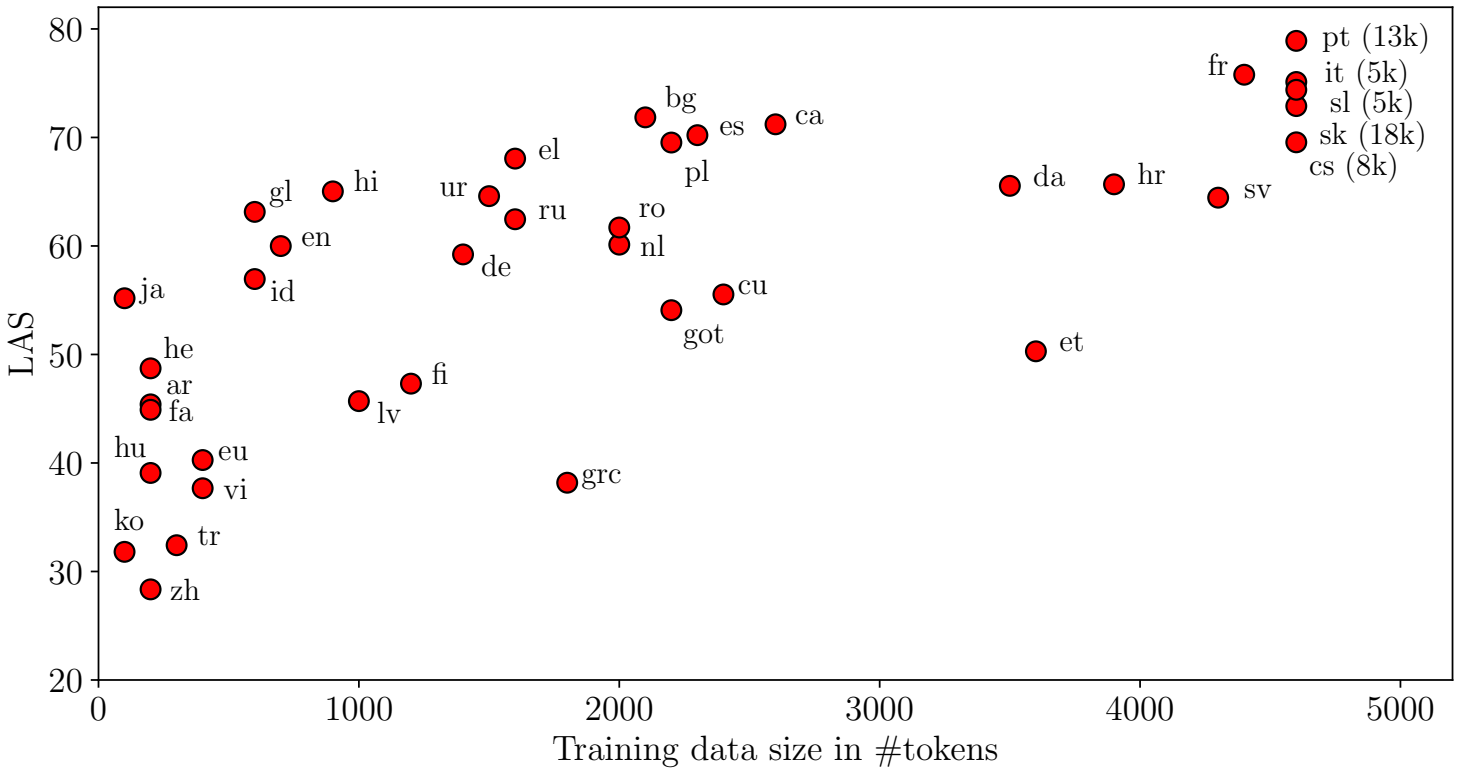
- LEX outperforms when there is only 1100 tokens (avg. 55 sentences)

Average results across 41 targets



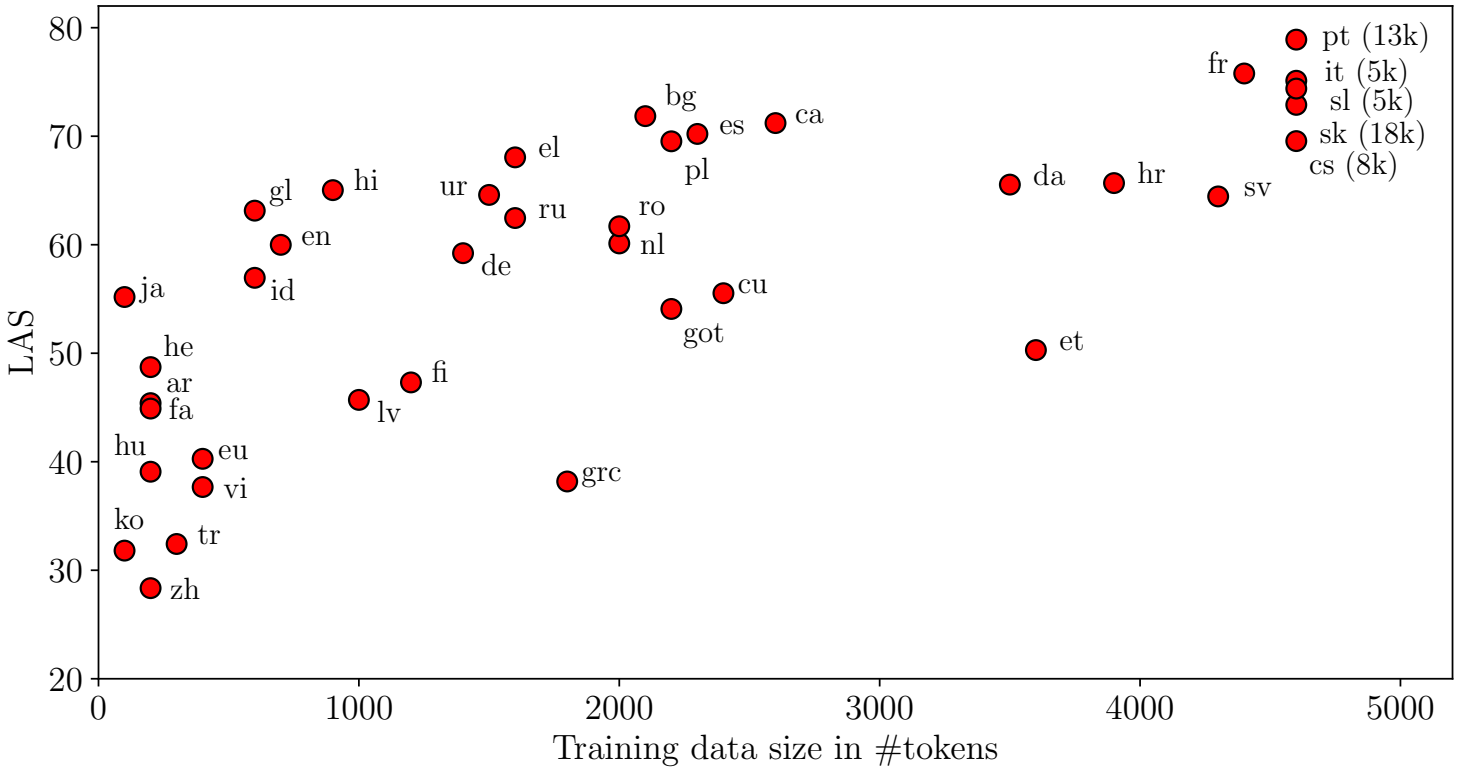
- Does it mean that if you have more than 55 sentences DELEX does not have sense?

Cutting points for all targets



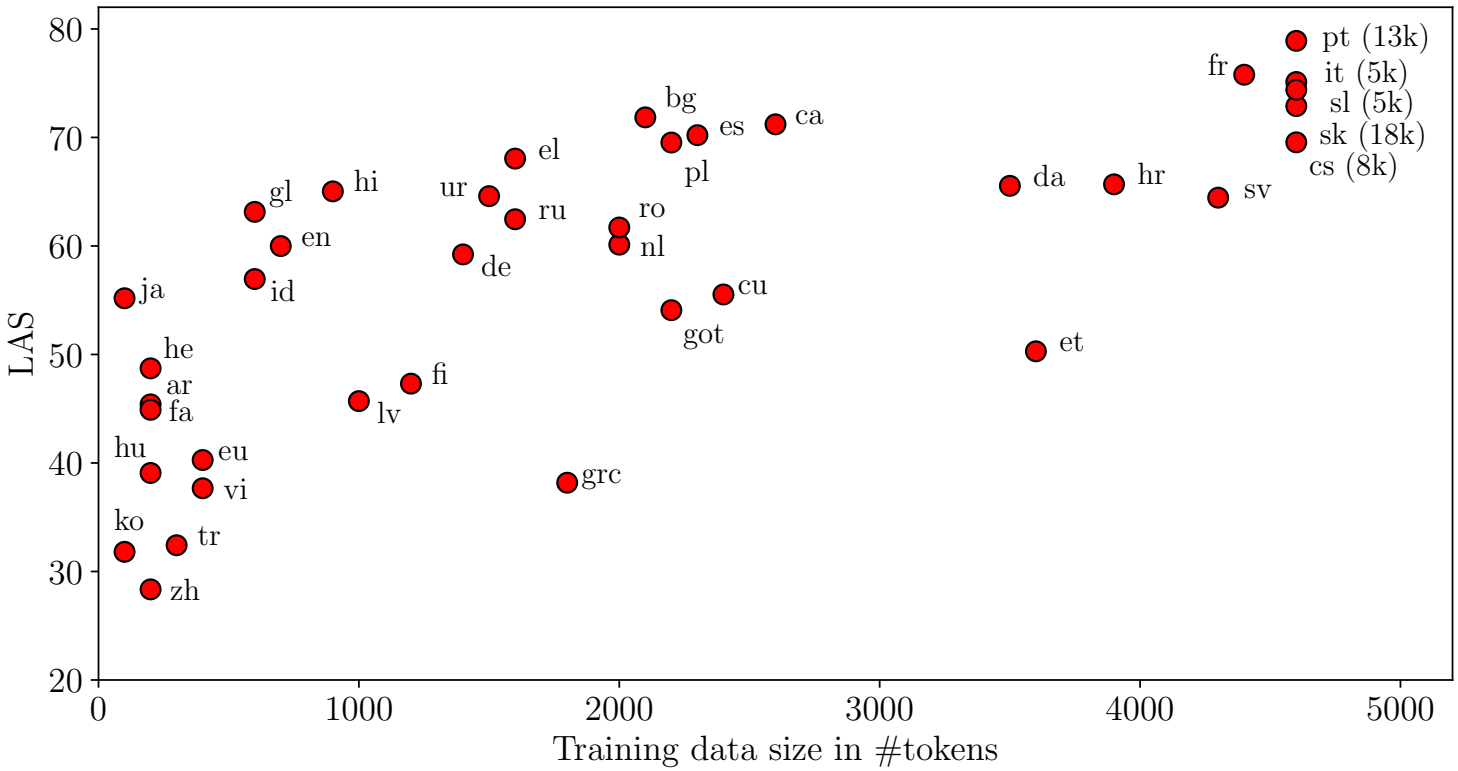
- Almost no language falls around 1100 tokens.

Cutting points for all targets



- For some LEX is always the best choice (even trained on 100 tokens!).

Cutting points for all targets

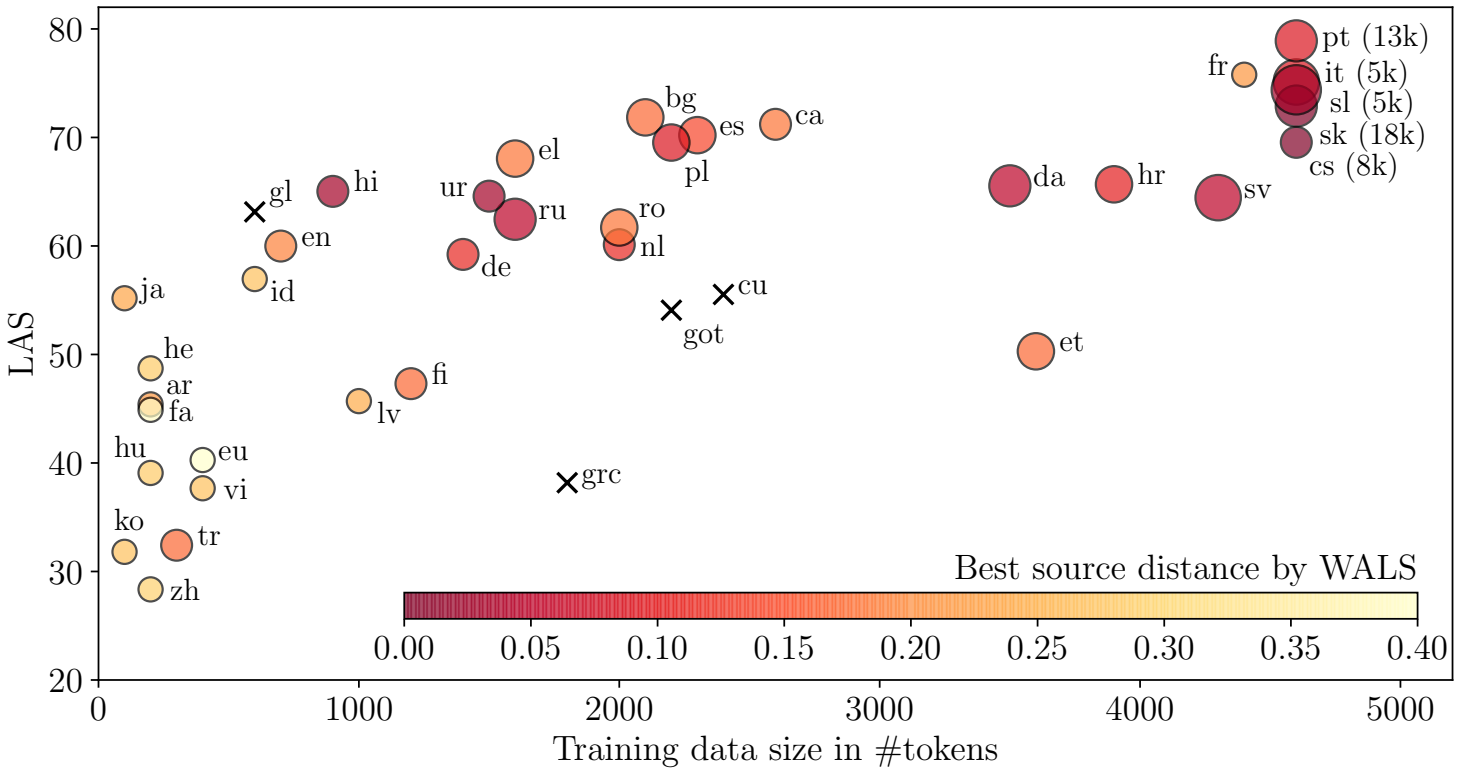


- **Conclusion**: when you have more tokens than 20k – use LEX.

How to analyze the cutting points?

- we want to analyze the influence of source languages
- we use The World Atlas of Language Structures (WALS)
  - how similar is the best source – color
  - how many good sources – size

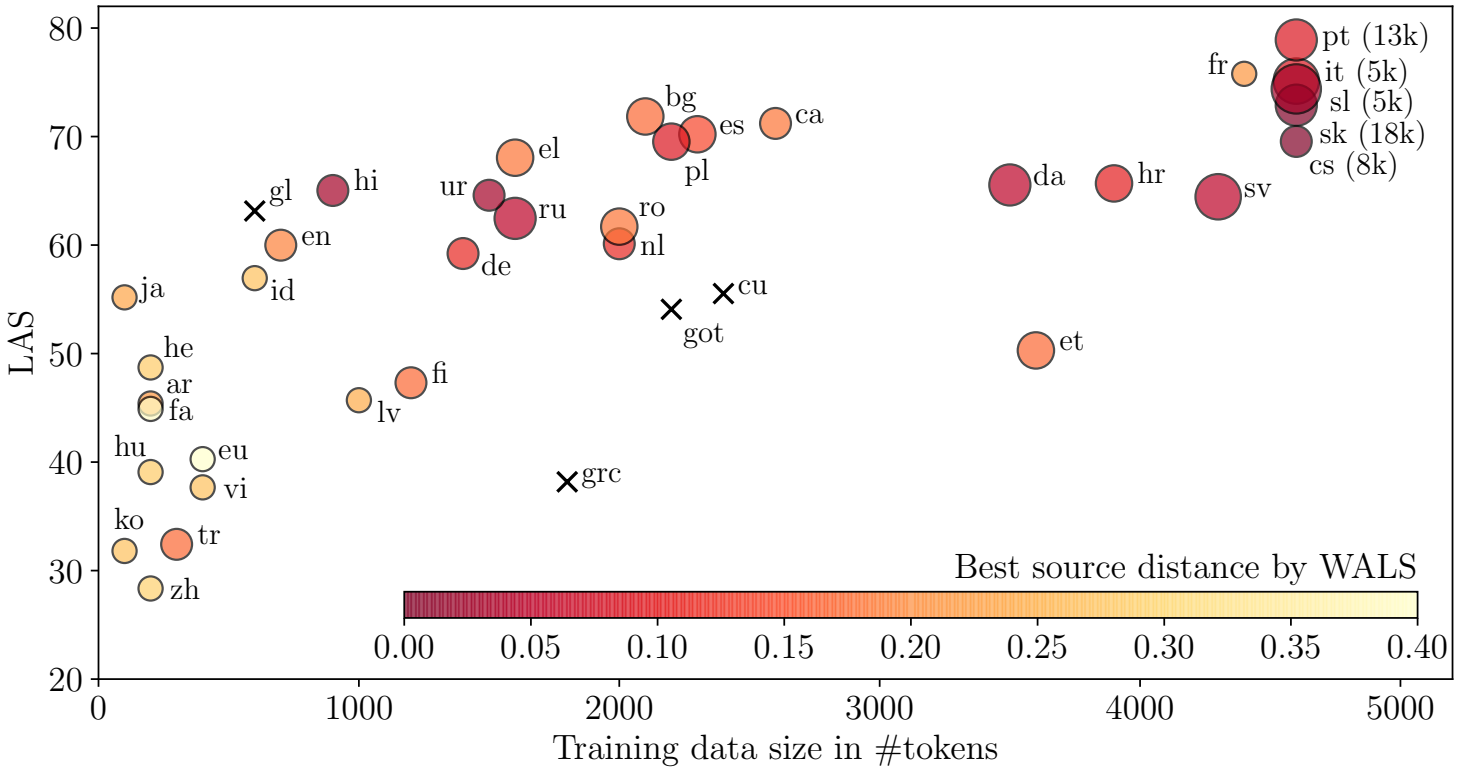
Cutting points with source strength



- The bigger and darker the point the better sources exist.

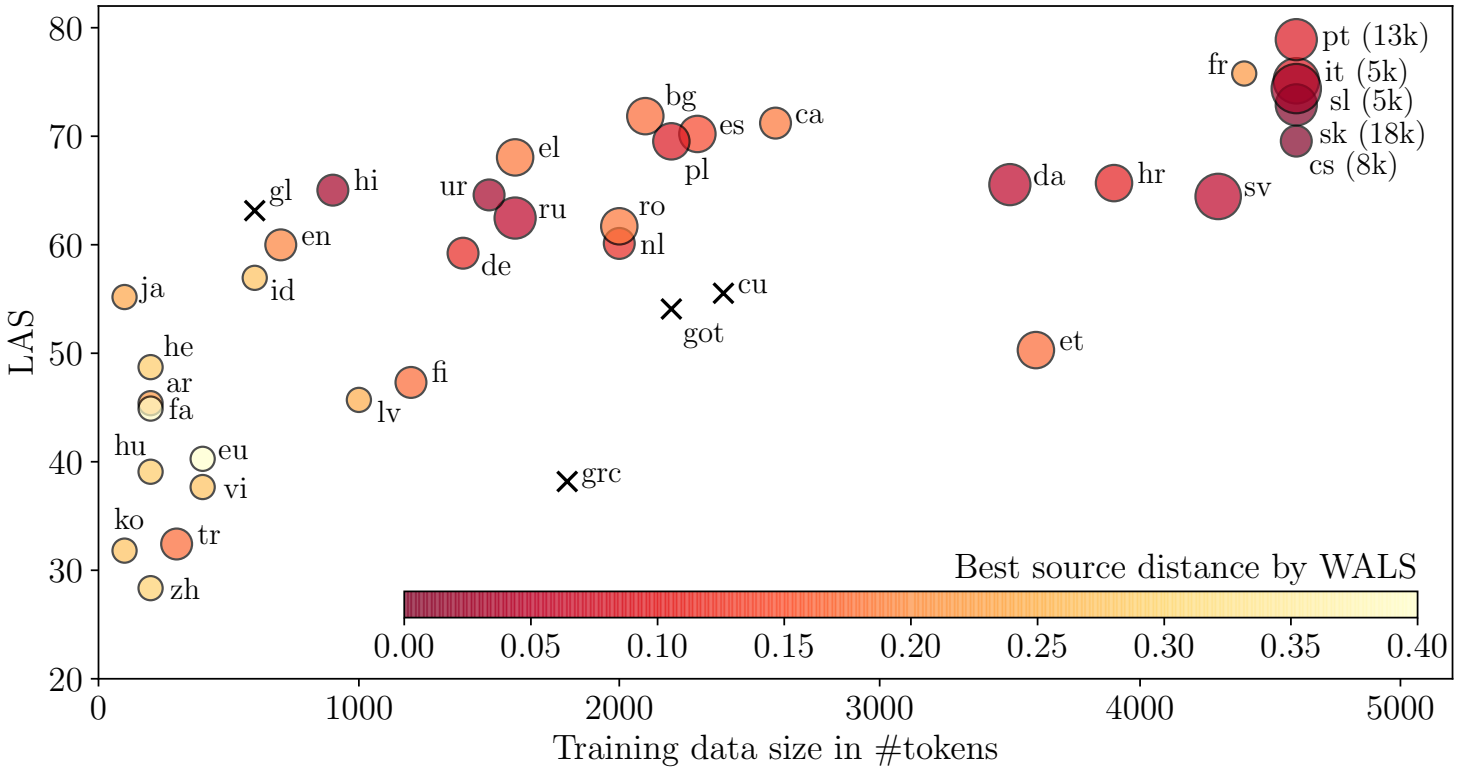
Cutting points with source strength

ExternalPOS



- **Conclusion**: when target belongs to an underrepresented language family – use LEX.

Cutting points with source strength

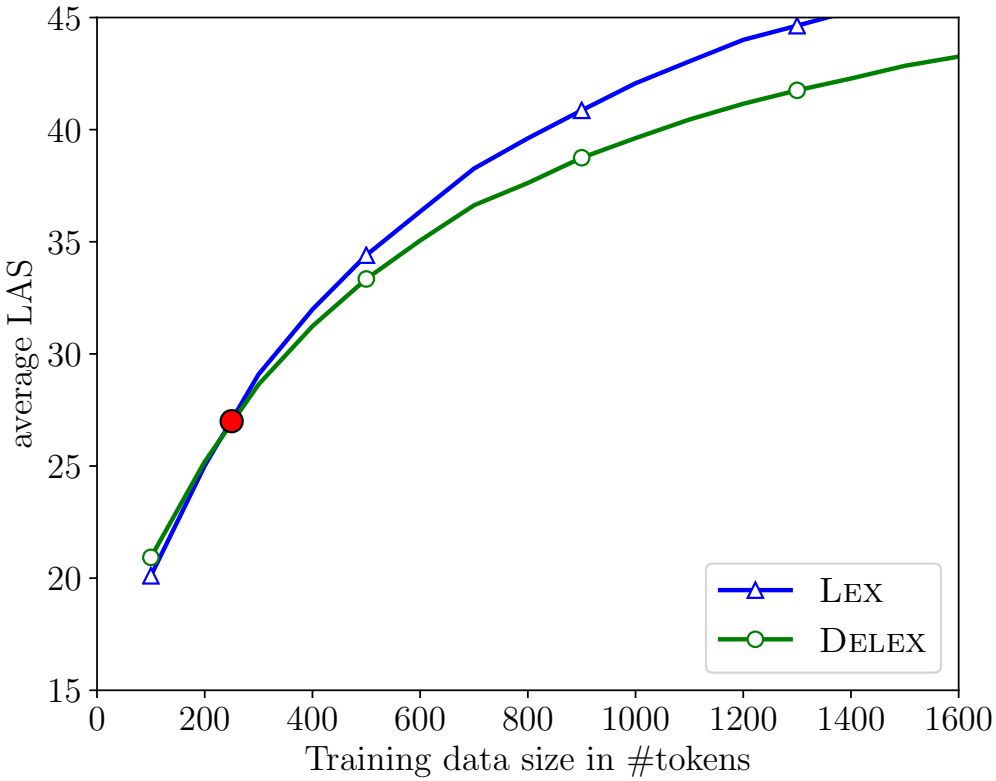


- **Conclusion**: when there is less trees than 20k but very good sources – use DELEX.

Table of Contents

Average results across 41 targets

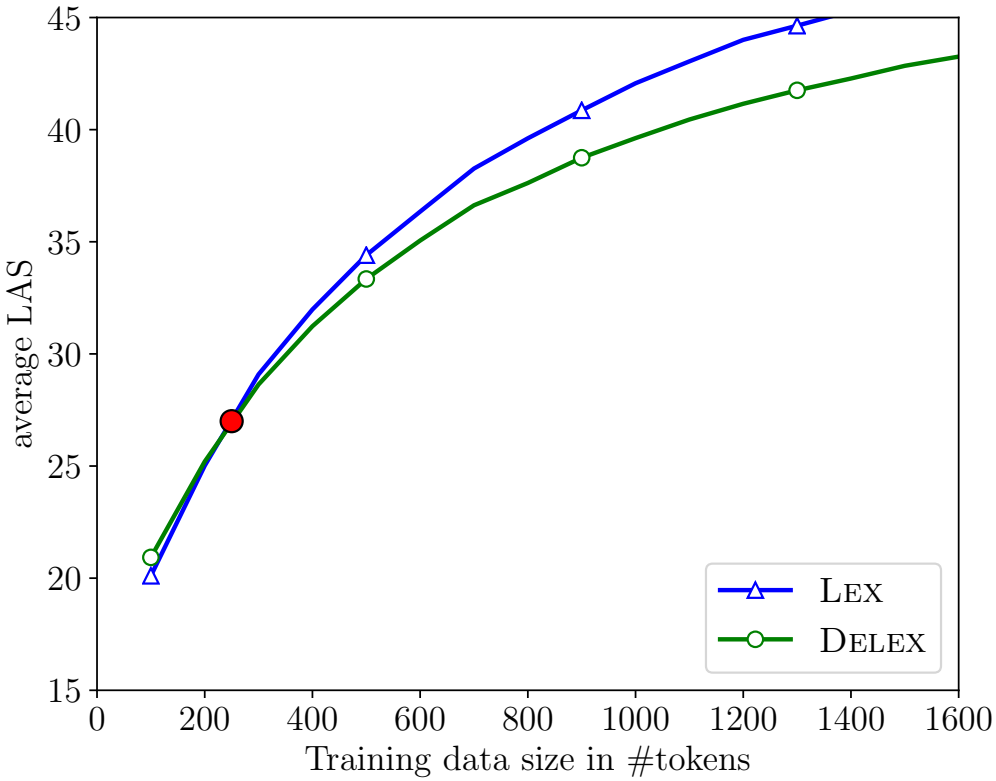
TreebankPOS



- The accuracy of parsers drops due to lower POS accuracy.

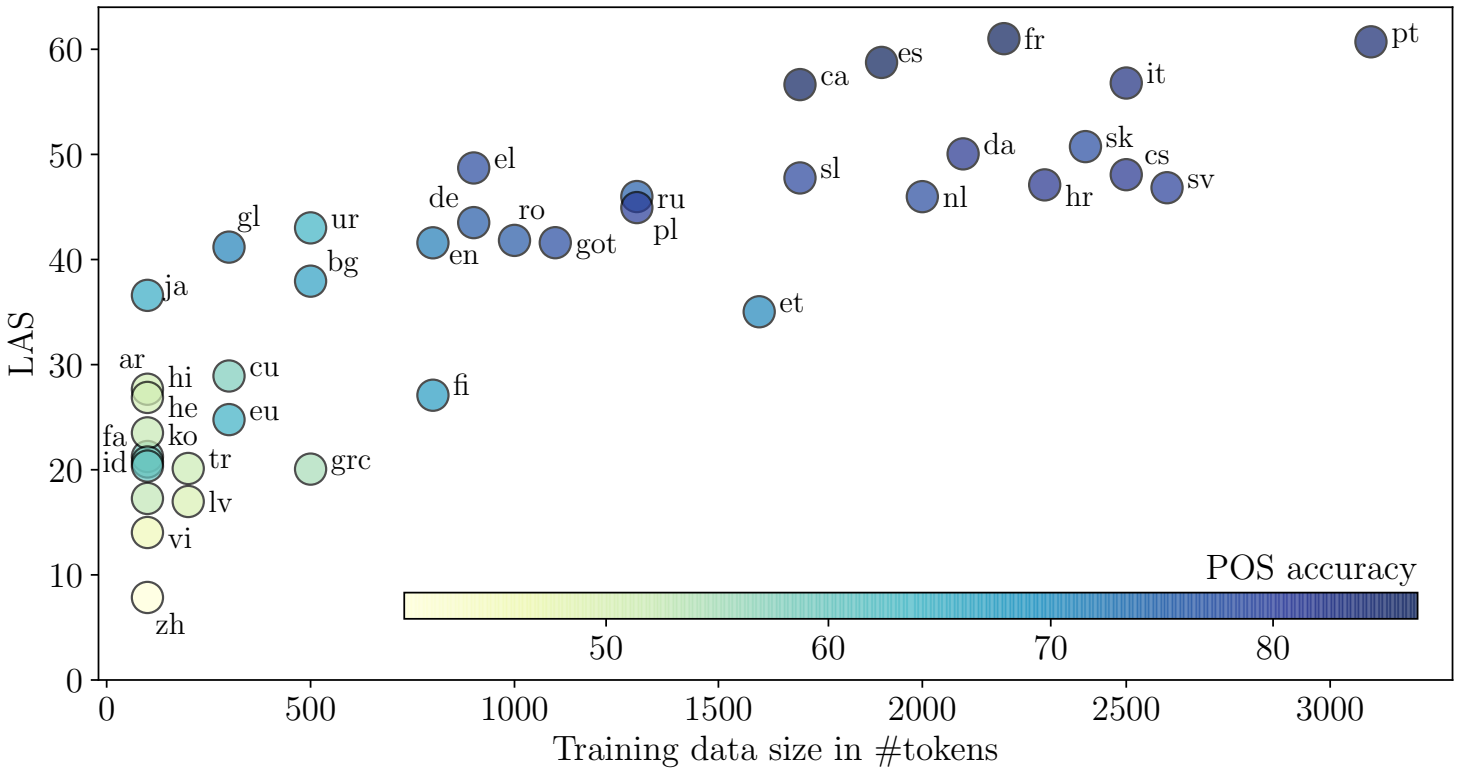
Average results across 41 targets

TreebankPOS



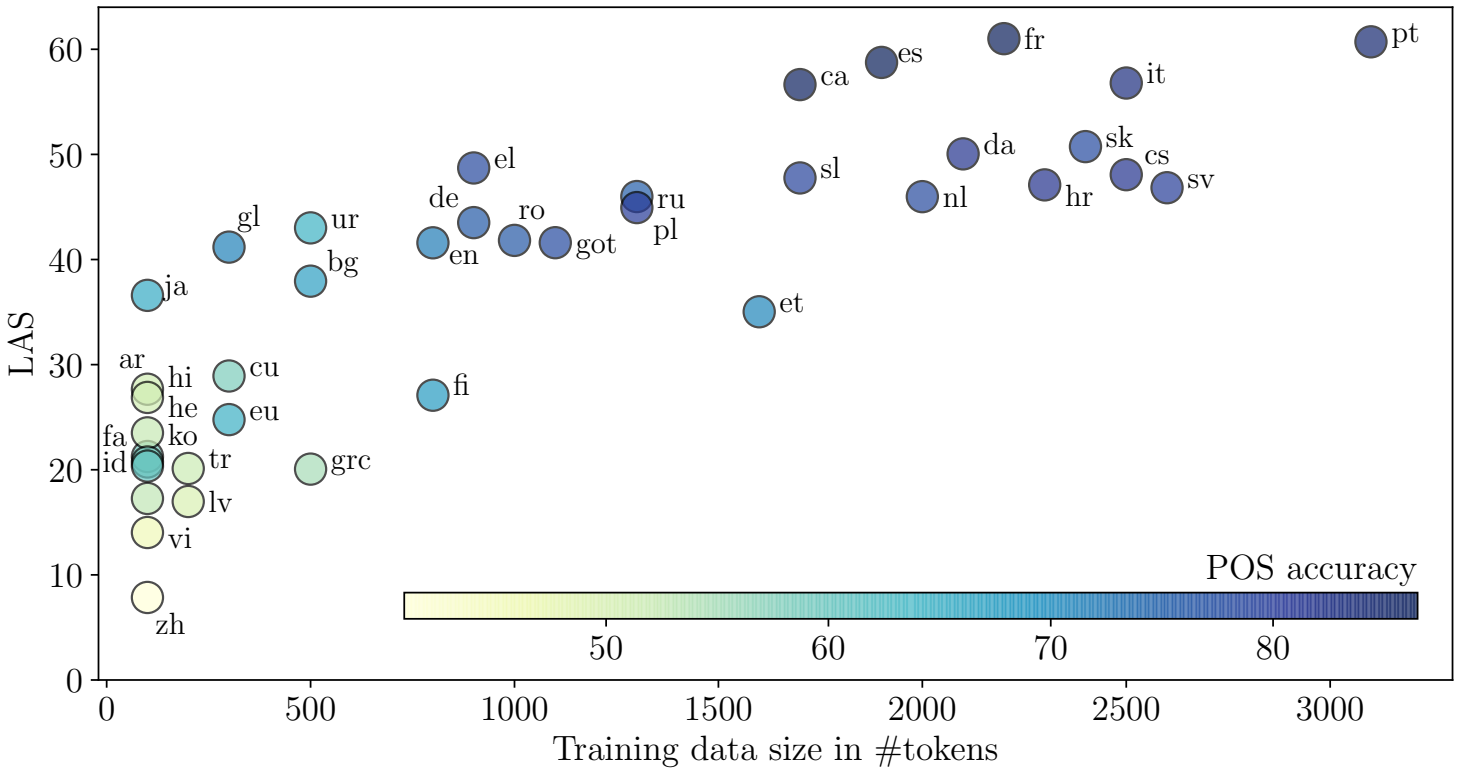
- LEX outperforms DELEX at only 250 tokens.

Cutting points for all targets



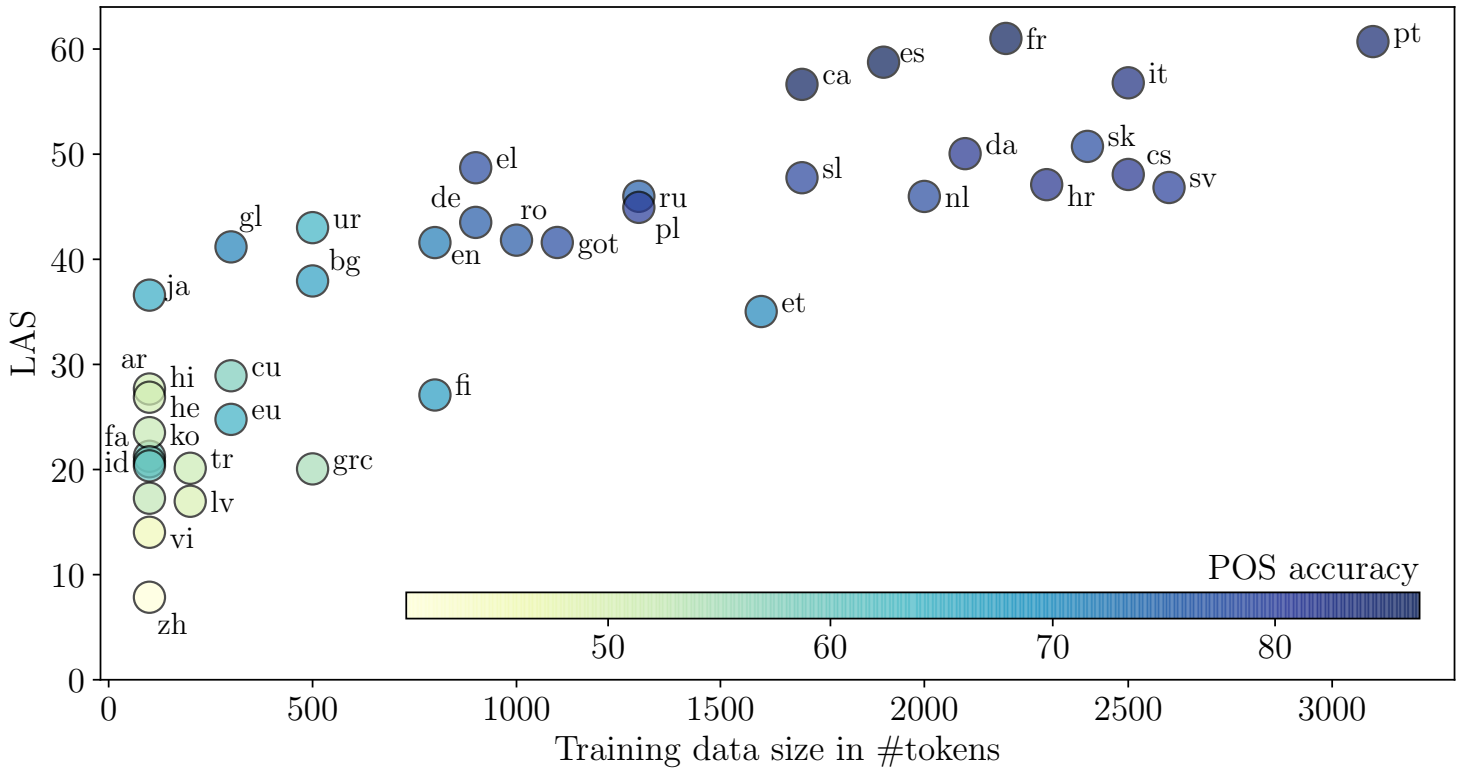
- The color indicates POS tagging accuracy.

Cutting points for all targets



- **Conclusion**: when you have more tokens than 3k – use LEX.

Cutting points for all targets



- **Conclusion**: when you have more than 1k tokens:
  - POS accuracy is reasonable (more than 70%)
  - then if you have good sources: use DELEX

Table of Contents

Intuition about the real test languages – ExternalPOS

Rules (conclusions from the plots):

- more tokens than 20k – use lex
- target belongs to an underrepresented family – use lex
- there is less trees than 20k but very good sources – use delex

	lex	delex
Buryat (19 sent.)		
Kurmanji (20 sent.)		
North Sami (20 sent.)		
Upper Sorbian (23 sent.)		

Intuition about the real test languages – ExternalPOS

Rules (conclusions from the plots):

- more tokens than 20k – use LEX
- target belongs to an underrepresented family – use LEX
- there is less trees than 20k but very good sources – use DELEX

	LEX	DELEX
Buryat (19 sent.)	?	?
Kurmanji (20 sent.)		
North Sami (20 sent.)		
Upper Sorbian (23 sent.)		

→ Buryat – Mongolic language without any close relatives among the source languages.

Intuition about the real test languages – ExternalPOS

Rules (conclusions from the plots):

- more tokens than 20k – use LEX
- **target belongs to an underrepresented family – use lex**
- there is less trees than 20k but very good sources – use DELEX

	LEX	DELEX
Buryat (19 sent.)	✓	×
Kurmanji (20 sent.)		
North Sami (20 sent.)		
Upper Sorbian (23 sent.)		

→ Buryat – Mongolic language without any close relatives among the source languages.

Intuition about the real test languages – ExternalPOS

Rules (conclusions from the plots):

- more tokens than 20k – use LEX
- target belongs to an underrepresented family – use LEX
- there is less trees than 20k but very good sources – use DELEX

	LEX	DELEX
Buryat (19 sent.)	✓	×
Kurmanji (20 sent.)	✓	×
North Sami (20 sent.)	×	✓
Upper Sorbian (23 sent.)	×	✓

Intuition about the real test languages – TreebankPOS

Rules (conclusions from the plots):

- more tokens than 3k – use LEX
- more than 1k tokens and very good sources – use DELEX

	LEX	DELEX
Latin (18k tokens)	✓	×
Irish (13k tokens)	✓	×
Ukrainian (12k tokens)	✓	×
Kazakh (529 tokens)	✓	×
Uyghur (1,5k tokens)	✓	×

Application to real target languages

		LEX	DELEX	LEX	DELEX
Surprise	Buryat	✓	×		
	Kurmanji	✓	×		
	North Sami	×	✓		
	Upper Sorbian	×	✓		
Small	Latin	✓	×		
	Irish	✓	×		
	Ukrainian	✓	×		
	Kazakh	✓	×		
	Uyghur	✓	×		

Application to real target languages

		LEX	DELEX	LEX	DELEX
Surprise	Buryat	✓	×	28.06	**32.01**
	Kurmanji	✓	×	**40.94**	38.17
	North Sami	×	✓	29.80	**35.80**
	Upper Sorbian	×	✓	49.59	**58.81**
Small	Latin	✓	×	**41.02**	35.06
	Irish	✓	×	**64.66**	44.54
	Ukrainian	✓	×	**64.81**	63.67
	Kazakh	✓	×	**26.55**	26.17
	Uyghur	✓	×	**34.81**	30.11

Application to real target languages

		LEX	DELEX	LEX	DELEX
Surprise	**Buryat**	✓	×	28.06	**32.01**
	Kurmanji	✓	×	**40.94**	38.17
	North Sami	×	✓	29.80	**35.80**
	Upper Sorbian	×	✓	49.59	**58.81**
Small	Latin	✓	×	**41.02**	35.06
	Irish	✓	×	**64.66**	44.54
	Ukrainian	✓	×	**64.81**	63.67
	Kazakh	✓	×	**26.55**	26.17
	Uyghur	✓	×	**34.81**	30.11

Table of Contents

Conclusions

Lexicalized vs. delexicalized?

- treebank size
- POS tagging accuracy
- typological relations between languages

We presented a methodology which worked for 8 out of 9 test languages.

Take home messages:

- Monolingual approach is a strong baseline.
- One should not be deceived by a small training treebank size.
- Never believe only your averages.

Conclusions

Lexicalized vs. delexicalized?

- treebank size
- POS tagging accuracy
- typological relations between languages

We presented a methodology which worked for 8 out of 9 test languages.

Take home messages:

- Monolingual approach is a strong baseline.
- One should not be deceived by a small training treebank size.
- Never believe only your averages.

Conclusions

Lexicalized vs. delexicalized?

- treebank size
- POS tagging accuracy
- typological relations between languages

We presented a methodology which worked for 8 out of 9 test languages.

Take home messages:

- Monolingual approach is a strong baseline.
- One should not be deceived by a small training treebank size.
- Never believe only your averages.

Conclusions

Lexicalized vs. delexicalized?

- treebank size
- POS tagging accuracy
- typological relations between languages

We presented a methodology which worked for 8 out of 9 test languages.

Take home messages:

- Monolingual approach is a strong baseline.
- One should not be deceived by a small training treebank size.
- Never believe only your averages.

Conclusions

Lexicalized vs. delexicalized?

- treebank size
- POS tagging accuracy
- typological relations between languages

We presented a methodology which worked for 8 out of 9 test languages.

Take home messages:

- Monolingual approach is a strong baseline.
- One should not be deceived by a small training treebank size.
- Never believe only your averages.

Conclusions

Lexicalized vs. delexicalized?

- treebank size
- POS tagging accuracy
- typological relations between languages

We presented a methodology which worked for 8 out of 9 test languages.

Take home messages:

- Monolingual approach is a strong baseline.
- One should not be deceived by a small training treebank size.
- Never believe only your averages.

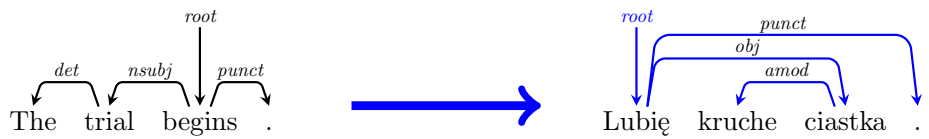
Table of Contents

Cross-lingual method – main idea

Delexicalized multisource model transfer and weighted blending.

Cross-lingual method – main idea

Delexicalized multisource **model transfer** and weighted blending.



Source: English

Target: Polish

Cross-lingual method – main idea

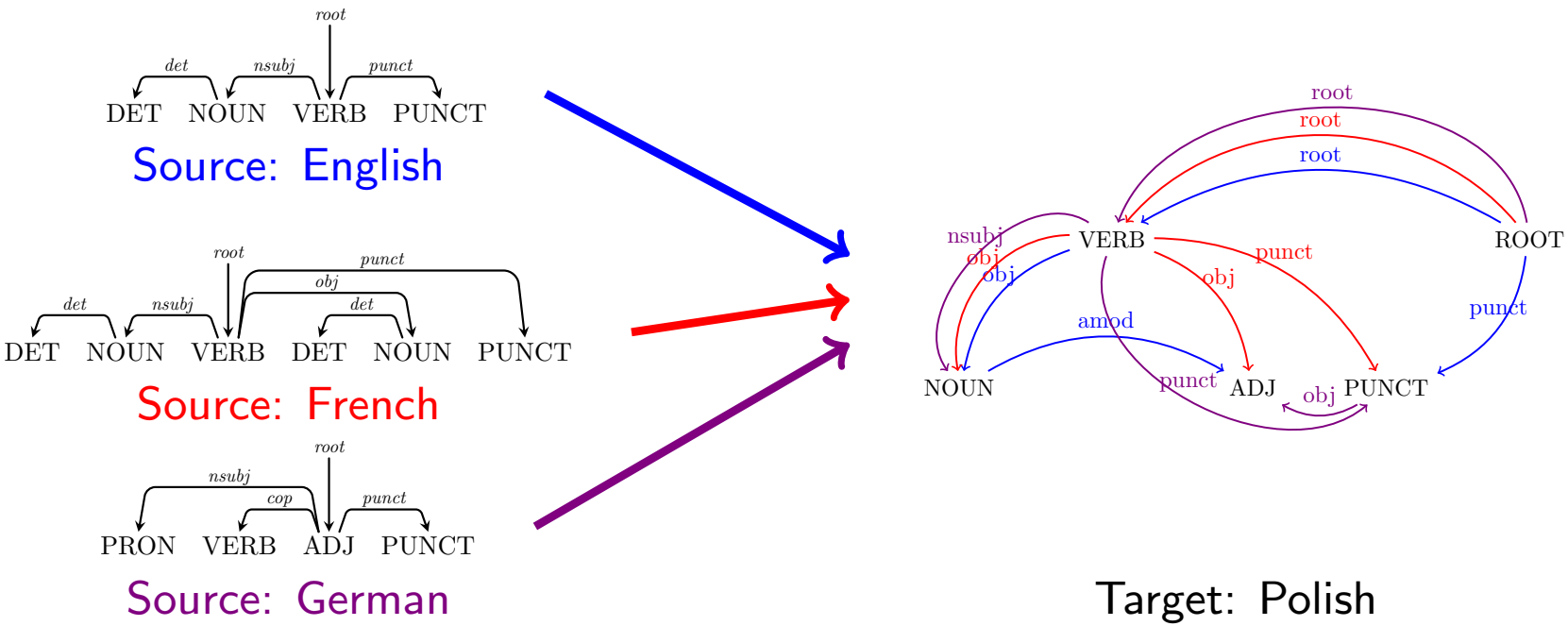
Delexicalized multisource model transfer and weighted blending.

Source: English

Target: Polish

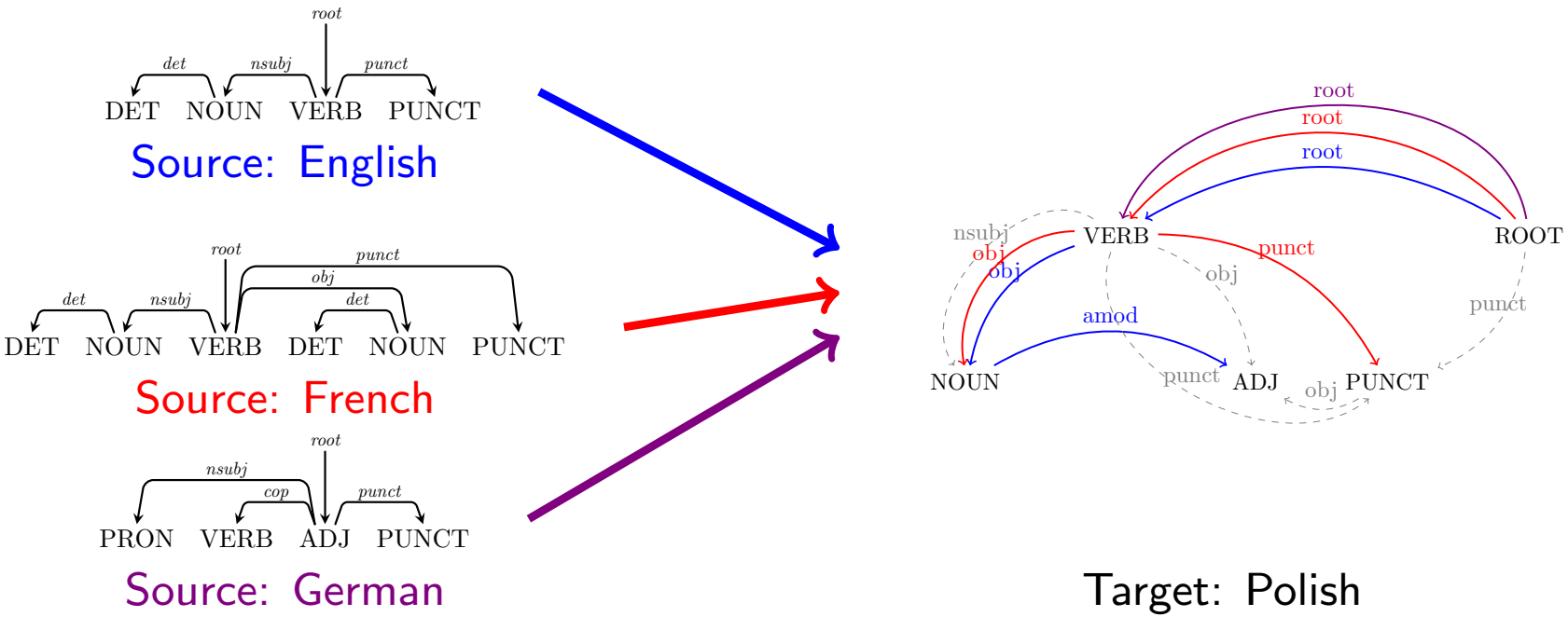
Cross-lingual method – main idea

Delexicalized **multisource** model transfer and weighted blending.



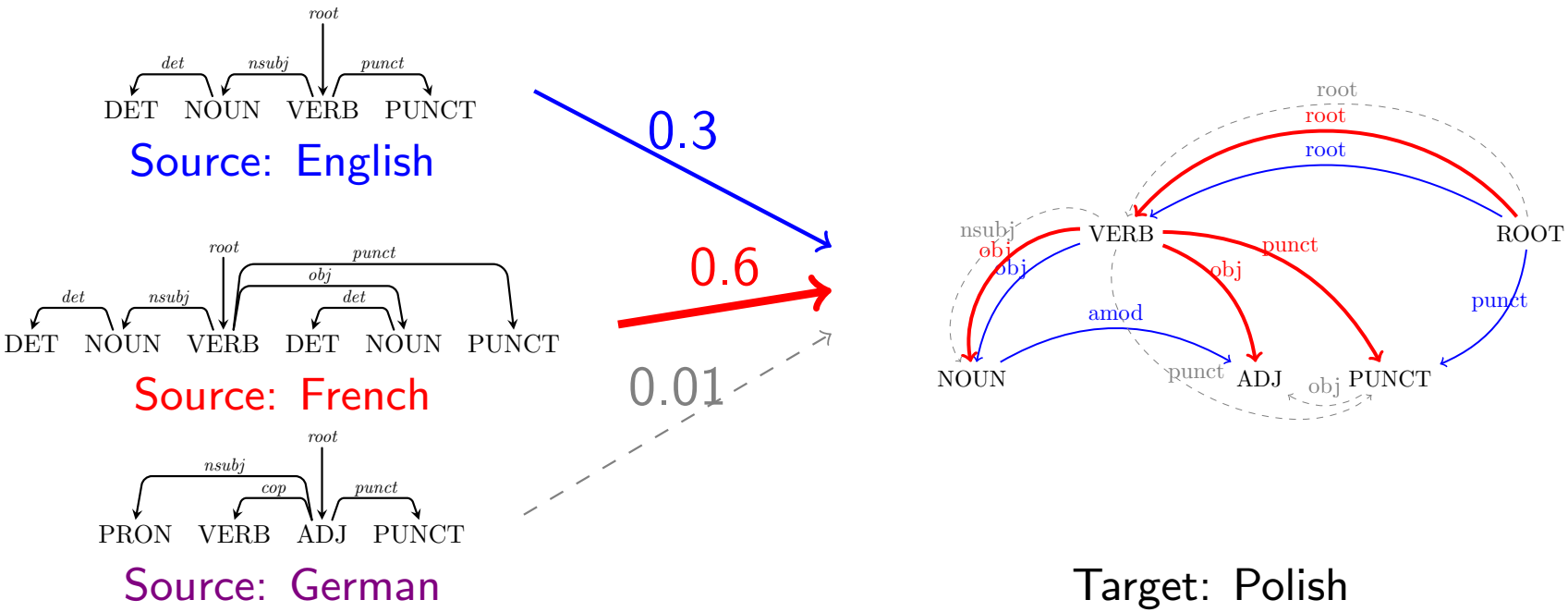
Cross-lingual method – main idea

Delexicalized multisource model transfer and weighted **blending**.



Cross-lingual method – main idea

Delexicalized multisource model transfer and **weighted** blending.



Cross lingual method – weighting methods

Source weighting methods::

- Gold POS tags (Rosa and Žabokrtský, 2015)
- Predicted POS tags, topological features (Agić, 2017)
- LAS on target trees samples

- Here only the last one – DELEX.

Cross lingual method – weighting methods

Source weighting methods::

- Gold POS tags (Rosa and Žabokrtský, 2015)
- Predicted POS tags, topological features (Agić, 2017)
- LAS on target trees samples

- Here only the last one – DELEX.

Cross lingual method – weighting methods

Source weighting methods::

- Gold POS tags (Rosa and Žabokrtský, 2015)
- Predicted POS tags, topological features (Agić, 2017)
- **LAS on target trees samples**

- Here only the last one – DELEX.

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*. Gothenburg Sweden, pages 1–10.

Anders Björkelund and Joakim Nivre. 2015. Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*. Association for Computational Linguistics, Bilbao, Spain, pages 76–86.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. Klcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of ACL-IJCNLP*.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *Proceedings of NAACL*. pages 129–132.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağr Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, pages 1–19.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*. Hyderabad, India.