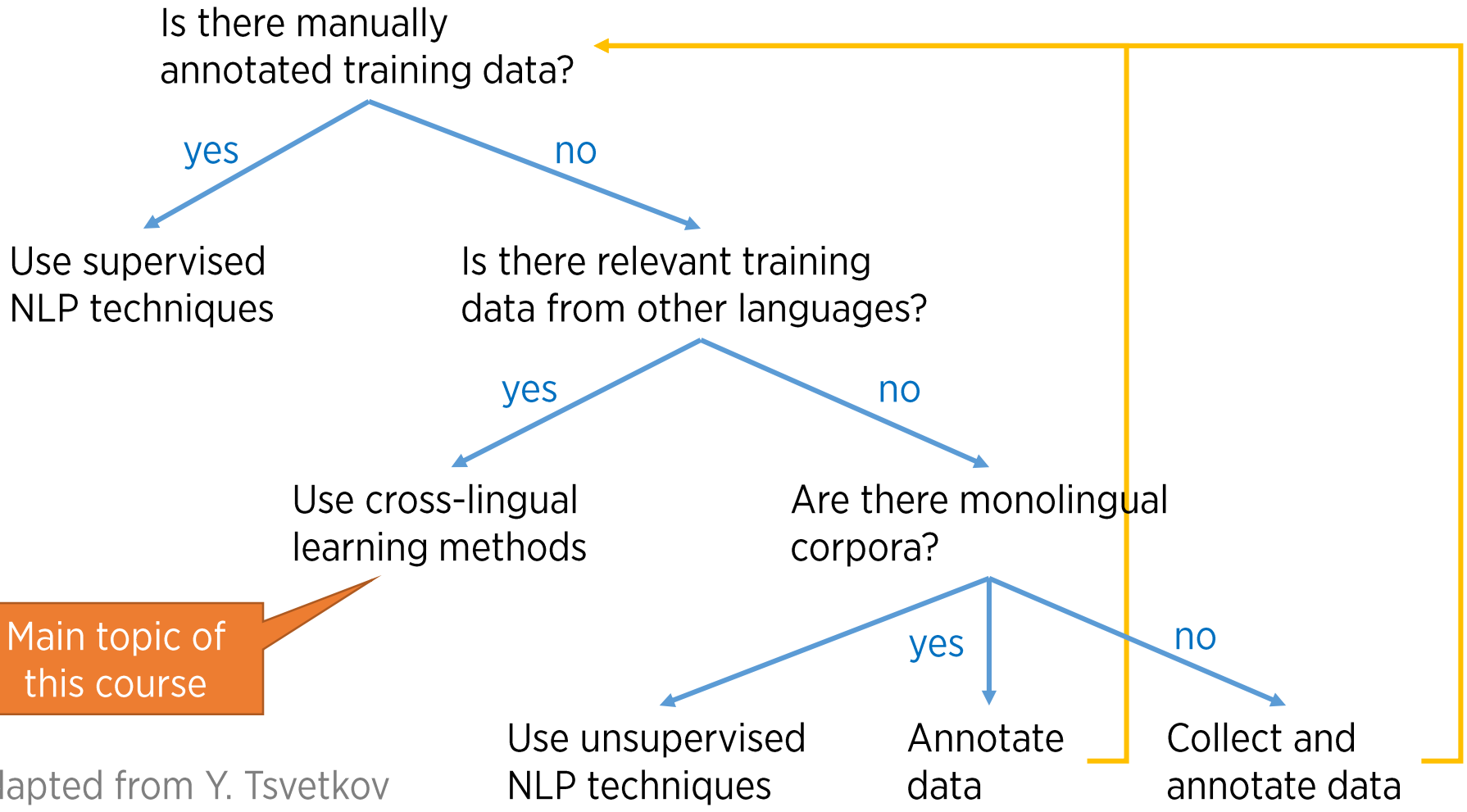
Language technology for low-resource languages

Day 3/5

LOT 2018, Groningen

Yves Scherrer, University of Helsinki

What resources can we get?



Adapted from Y. Tsvetkov

Data transfer approaches

- Annotation projection:
  - Build task model for language A
  - Apply word alignment tool
  - Build task model for language B
- Training data translation:
  - Build MT model for A – B
  - Build task model for language B
- Test data backtranslation:
  - Build task model for language A
  - Build MT model for B – A
- 2 models = 2 opportunities to produce errors…
  - Can we get by with a single model?

Cross-lingual learning techniques

- We have the data to train a model for language A, but need a model for language B
- Data transfer:
  - Convert the training data to language B
  - Train a model on the converted data of language B
  - Annotation projection, data translation
- Model transfer:
  - Train a model for language A
  - Convert the model so that it applies to language B
  - Plain model transfer, delexicalization, relexicalization
- Multilingual/multitask models

Model transfer

Overview

- Plain model transfer:
  - Pretend that languages A and B are in fact the same
  - Train a model on A, apply it to B without change
- Delexicalisation:
  - Train a model on A, remove all language-specific features (e.g. word forms) and replace them by language-independent features
- Relexicalisation:
  - Train a model on A, replace all A-specific features by B-specific features

Overview

- **Parallel corpora** are not strictly required for model transfer methods, but…
  - … they may be useful for finding language-independent features
  - … they may be useful for matching A-specific with B-specific features

Plain model transfer Tagging

- Don’t even try it for unrelated languages…

High-resource language (A)	“Low-resource” language (B)	Accuracy	Tagset size
Spanish	Catalan Aragonese	58% 72%	42
German	Dutch	24%	55
Czech	Slovak Polish	81% 66%	57
Slovak (+Translit.) Polish (+Translit.) Ukrainian	Rusyn	43% 50% 63%	657 920 1040

Scherrer, Y. (2014): *Unsupervised adaptation of supervised part-of-speech taggers for closely related languages*. Proceedings of the VarDial Workshop 2014.
Scherrer, Y. & Rabus, A. (2017): *Multi-source morphosyntactic tagging for Spoken Rusyn*. Proceedings of the VarDial Workshop 2017.

Plain model transfer Dependency parsing

- VarDial 2017 shared task on cross-lingual dependency parsing – Baselines:

High-resource language	"Low-resource" language	LAS	UAS
Slovenian	Croatian	53.35	63.94
(supervised)	Croatian	68.51	75.61
Danish	Norwegian	54.91	64.53
Swedish	Norwegian	56.63	66.24
(supervised)	Norwegian	78.23	82.28
Czech	Slovak	53.72	65.70
(supervised)	Slovak	69.14	76.57

M. Zampieri et al. (2017): *Findings of the VarDial Evaluation Campaign 2017*. Proceedings of the VarDial Workshop 2017.

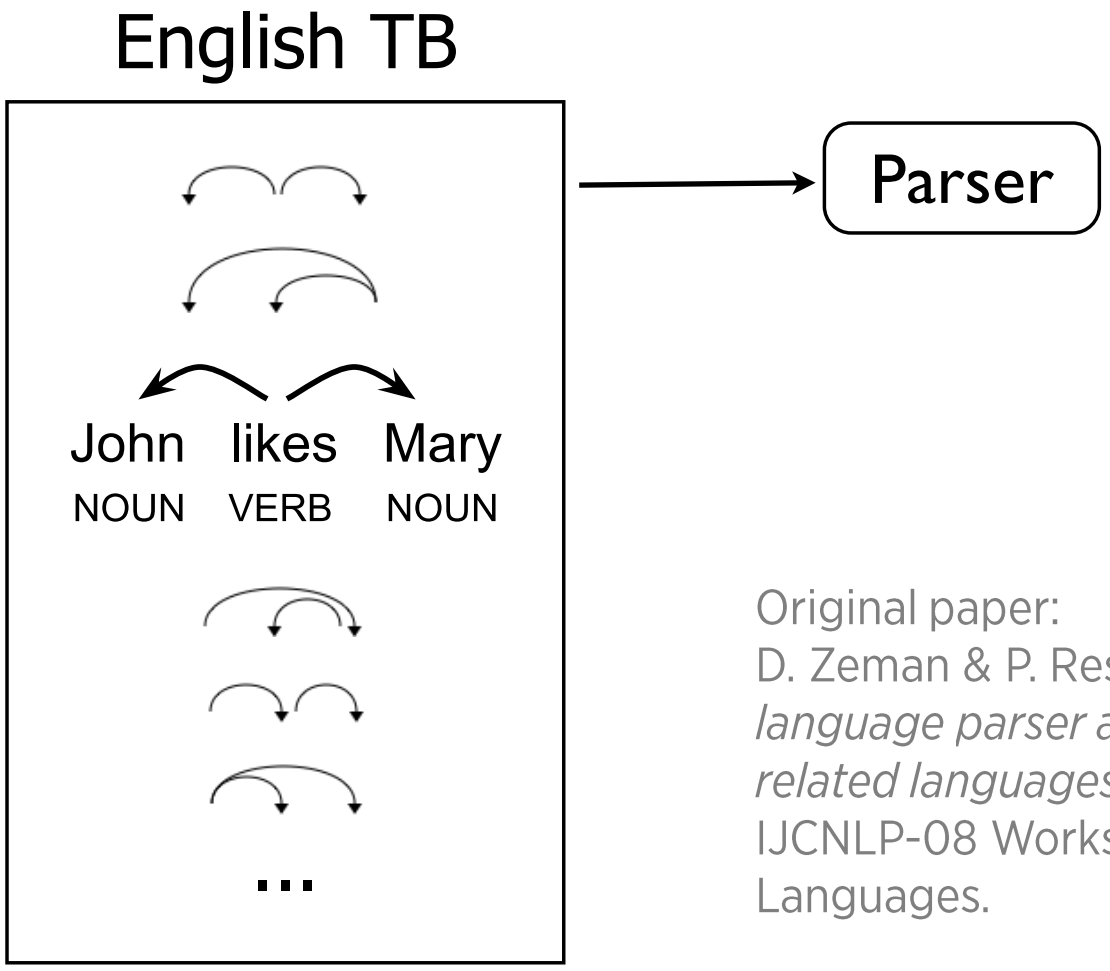
Plain model transfer

- Plain model transfer (kind of) works because some words happen to be the same in both languages (and because word order largely is the same)
  - Numbers, punctuation signs, proper nouns
  - Loanwords, cognates
- For unrelated languages, there are few such cues
- Idea: use language-independent cues
  - For a parser, part-of-speech tags (in a common format) are language-independent cues

Reading

- Ryan McDonald, Slav Petrov & Keith Hall (2011): *Multi-source transfer of delexicalized dependency parsers*. Proceedings of EMNLP. https://www.aclweb.org/anthology/D11-1006
- Questions:
  - What is delexicalization?
  - The authors propose to use delexicalization for parsing. Would this approach also work for other tasks, such as part-of-speech tagging?

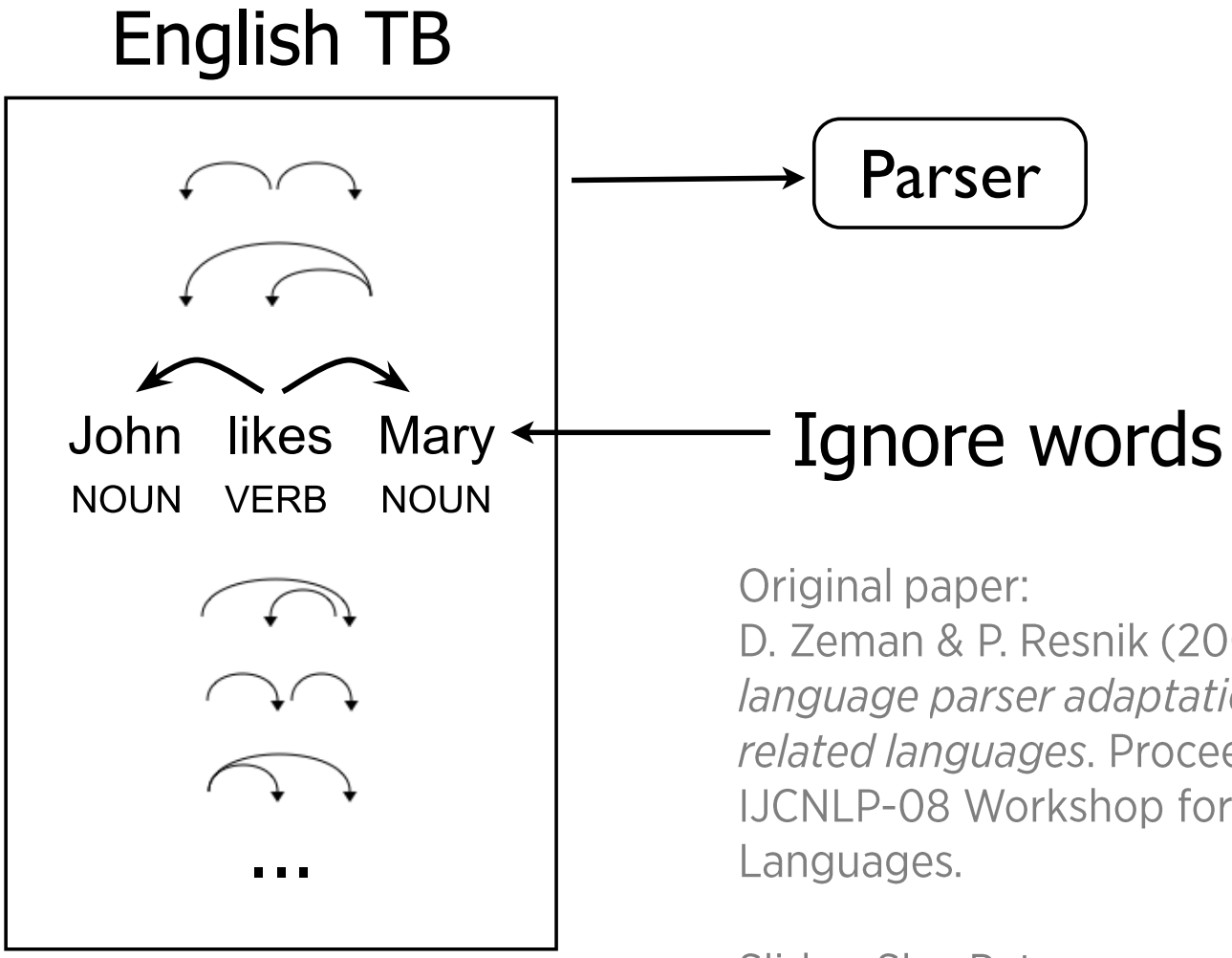
Delexicalization – Parsing



Original paper:
D. Zeman & P. Resnik (2008): *Cross-language parser adaptation between related languages*. Proceedings of the IJCNLP-08 Workshop for Less Privileged Languages.

Slides: Slav Petrov

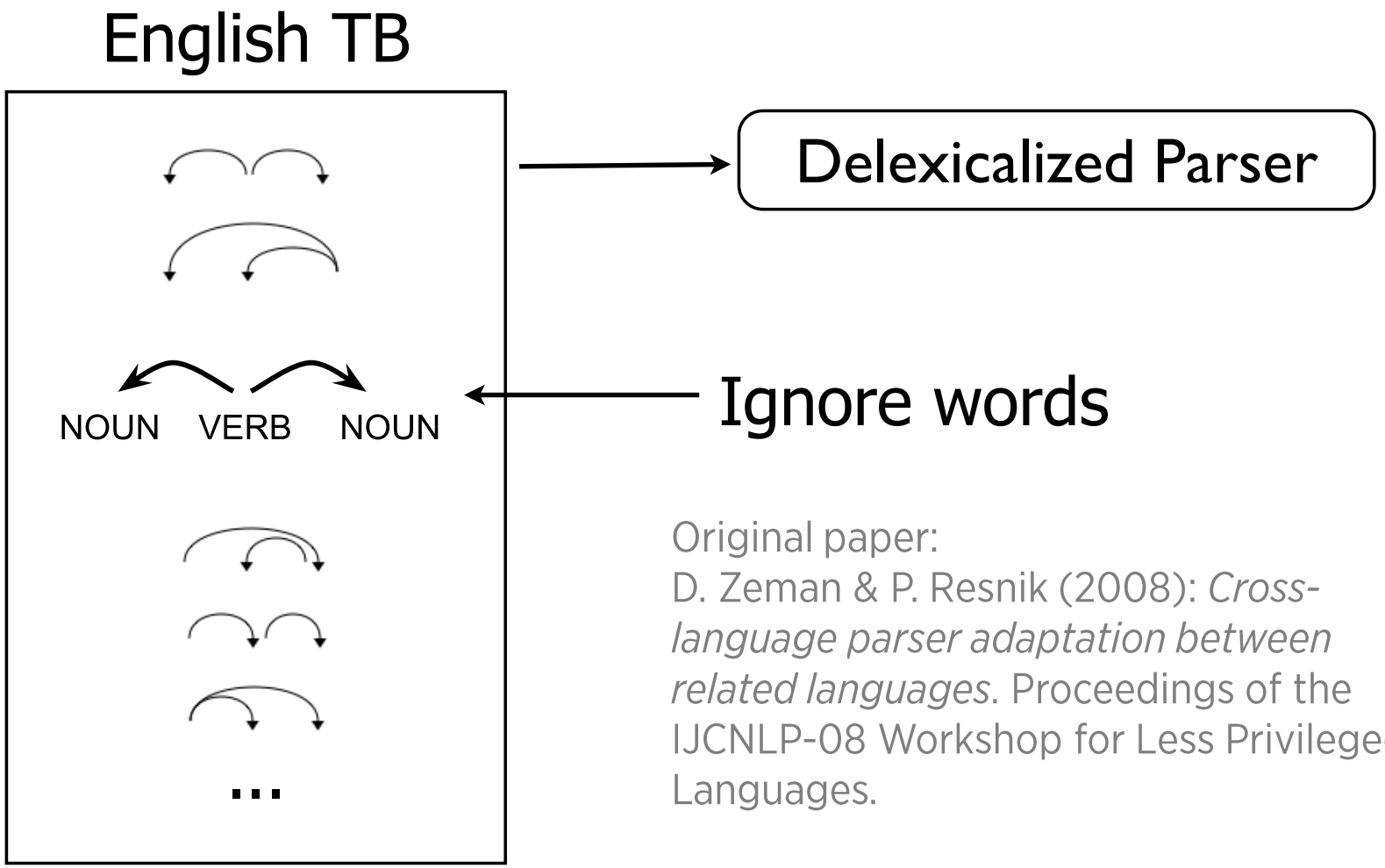
Delexicalized Parsing



Original paper:
D. Zeman & P. Resnik (2008): *Cross-language parser adaptation between related languages*. Proceedings of the IJCNLP-08 Workshop for Less Privileged Languages.

Slides: Slav Petrov

Delexicalized Parsing

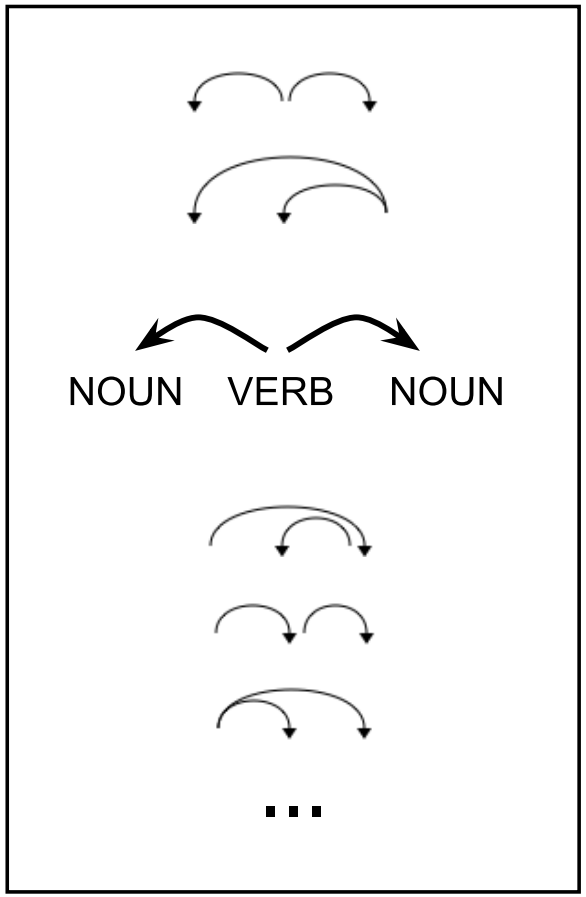


Original paper:
D. Zeman & P. Resnik (2008): *Cross-language parser adaptation between related languages*. Proceedings of the IJCNLP-08 Workshop for Less Privileged Languages.

Slides: Slav Petrov

Delexicalized Parsing

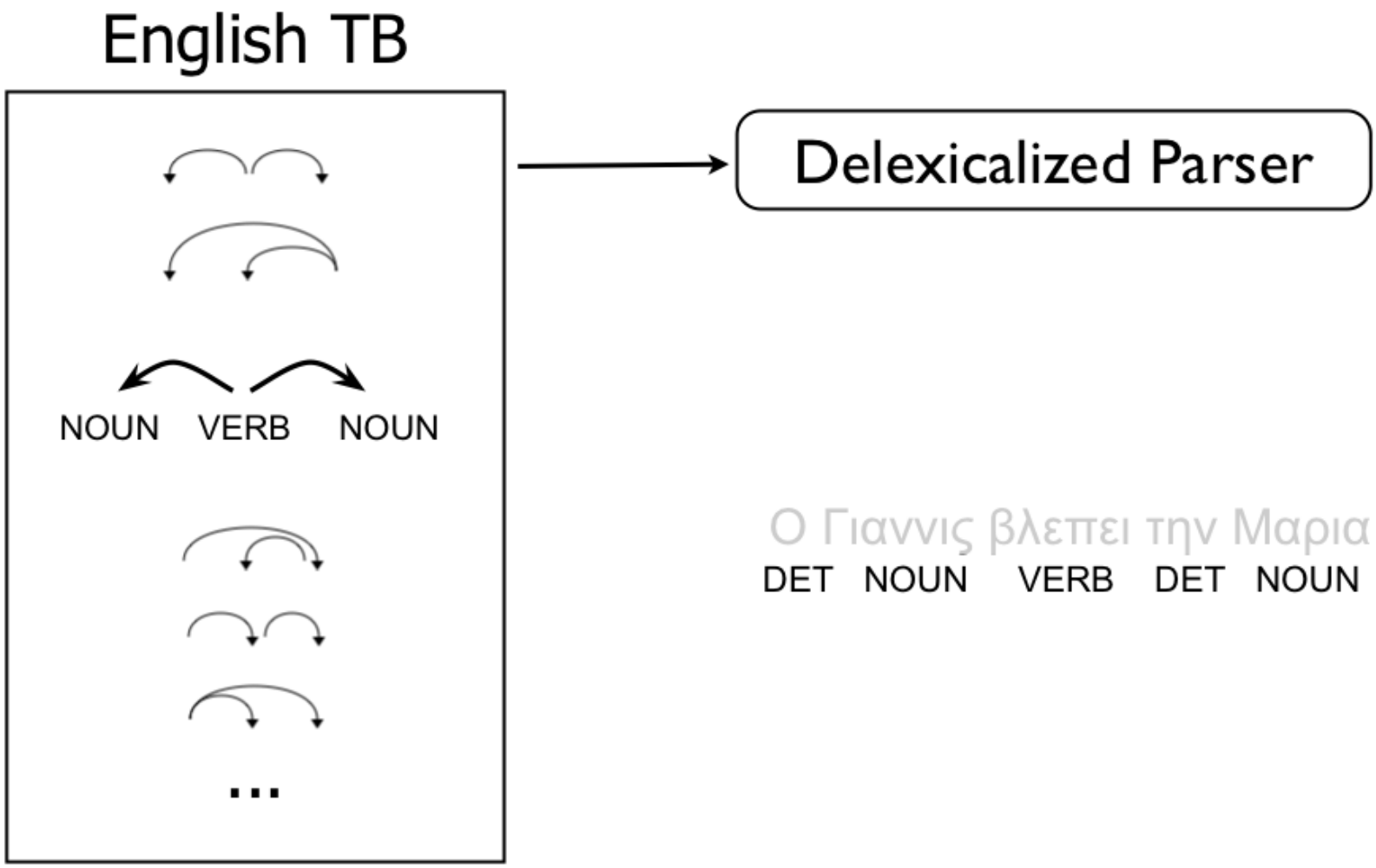
English TB



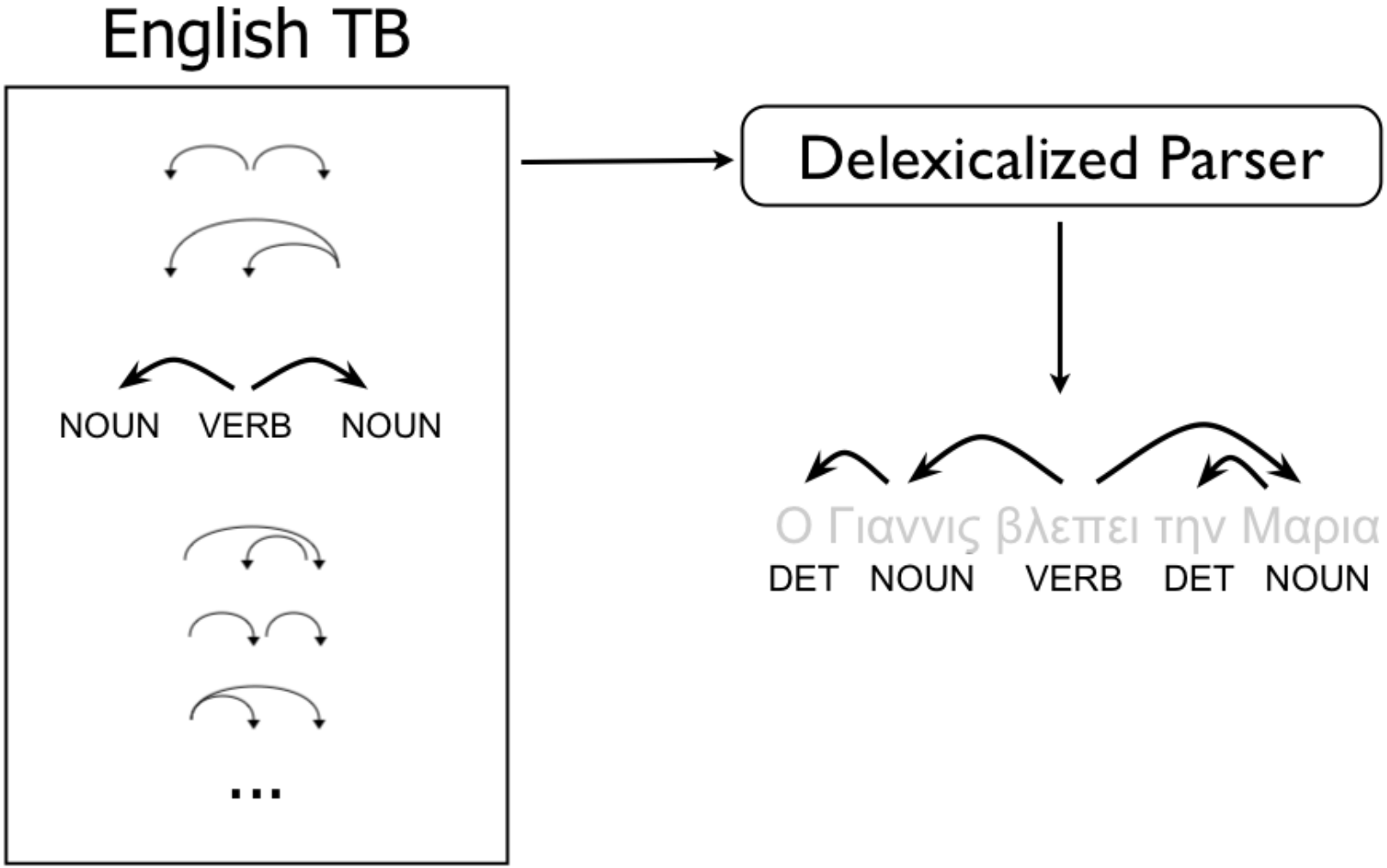
Delexicalized Parser

Ο Γιαννις βλεπει την Μαρια
DET NOUN VERB DET NOUN

Delexicalized Parsing



Delexicalized Parsing



Delexicalized Parsing

- Does this work?
  - Lexicalized parser for English (standard): 89% UAS
  - Delexicalized parser for English: 83% UAS
  - The POS tag is the most important information for a parser.

- Original work is on Danish -> Swedish:
  - Train a Danish parser and apply it to Swedish without change (plain model transfer): 43.28% F-measure
  - Train a delexicalized parser on Danish and apply it to Swedish: 65.50% F-measure
  - Train a lexicalized parser on Swedish (high-resource setting): 77.81% F-measure

Delexicalized Parsing

- The languages need to have similar word order
- Example: N ADJ N
  - EN: Adjectives are prenominal, so the parser learns a dependency from the right N to ADJ
  - FR: Adjectives are (mostly) postnominal, so a parser trained on English will predict a wrong dependency from the right N
- Delexicalized parsing assumes that part-of-speech taggers are available for both languages

Delexicalization

- Common parsing models look at several features to create the parse tree: word form, lemma, POS tag, ...
- Delexicalization deprives the parsing model of some features while retaining others.
  - Language-dependent features (word form, ...) are removed
  - Language-independent features (POS tag, ...) are retained
- If delexicalization works for parsing, could we do the same for tagging?
  - What language-dependent features are useful for tagging?
  - What language-independent features are useful for tagging?

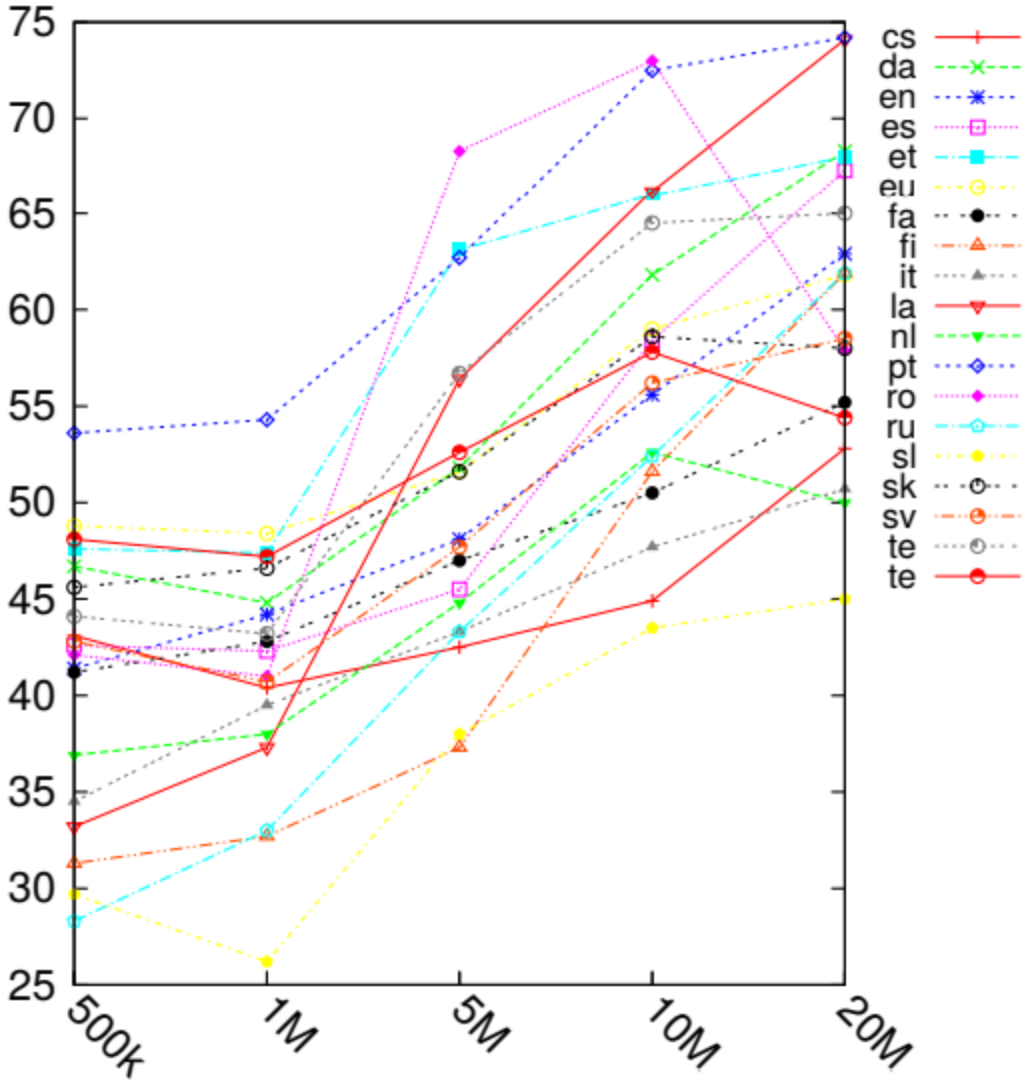
Delexicalized tagging

- What language-dependent features are useful for tagging?
  - Word forms
  - If we remove them, there is nothing left ☹

- What language-independent features are useful for tagging?
  - Tag sequences (assumed to be identical as in source language)
  - Word shape (punctuation, numbers, suffixes, word length…)
  - Word frequency
  - Previous and following words and how well they predict the current word
  - Sentence length

- Z. Yu, D. Mareček, Z. Žabokrtský, D. Zeman: If you even don't have a bit of Bible: Learning Delexicalized POS taggers. LREC 2016.

Delexicalized tagging



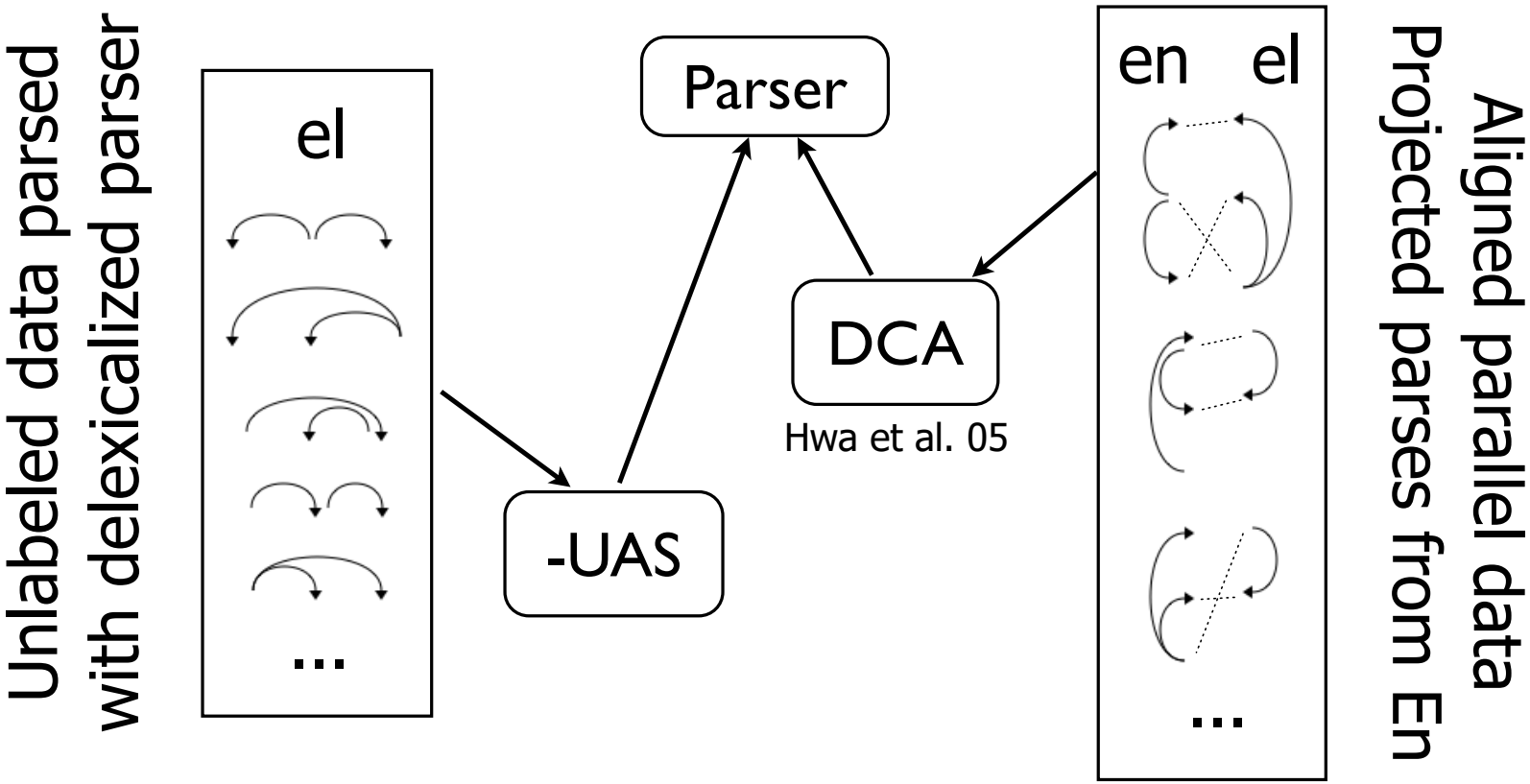
- Different sizes of data used to compute the word statistics
- Multi-source model with 7 source languages
- Does not work particularly well: 50%-60% for most languages

Delexicalization

- Delexicalization in itself rarely works well
  - Too much useful information is thrown away
  - Assumptions about language similarity are too strong

- But we can improve delexicalization in several ways:
  - If parallel data are available, we can combine delexicalization with annotation projection
  - We can replace the word forms by abstract word representations that are comparable across languages
  - If we have a bilingual dictionary, we can simply translate the word forms: *relexicalization*

Delexicalization + Annotation projection



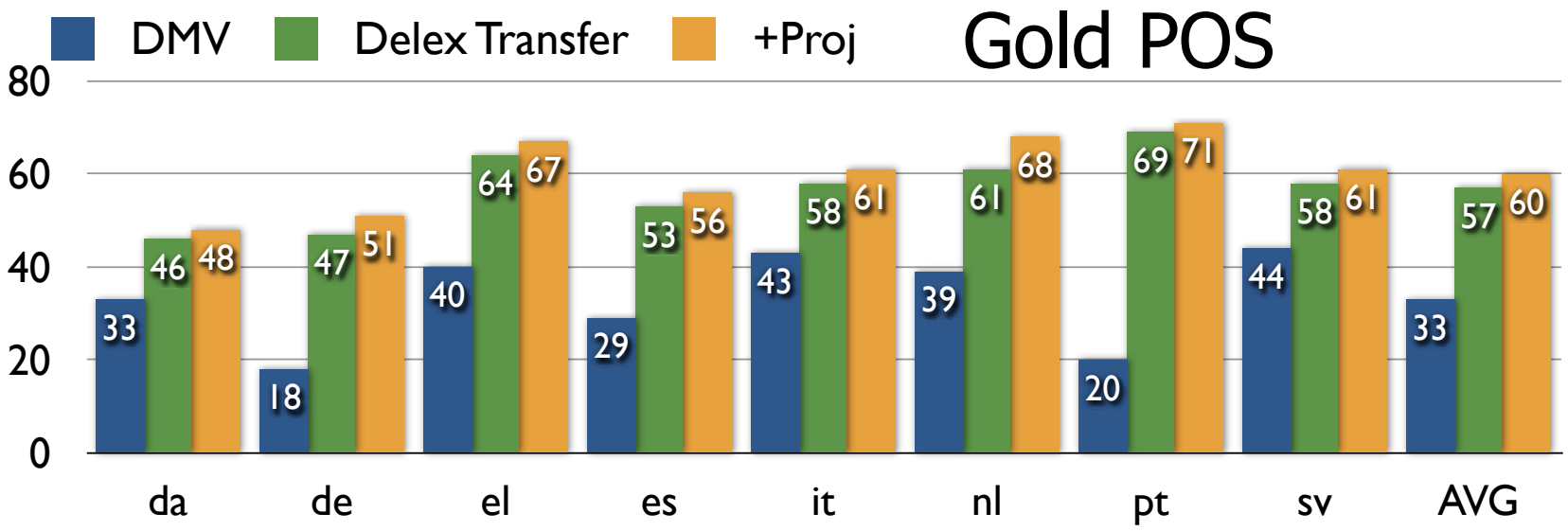
R. McDonald, S. Petrov & K. Hall (2011): Multi-source transfer of delexicalized dependency parsers. EMNLP 2011.

Slides by Slav Petrov

Delexicalization + Annotation projection

- Given: parallel corpus, aligned, parsed on the HRL side
- A part of the LRL side is parsed using a delexicalized parser
- This parsed corpus is then used to train a new LRL parser
  - This parser contains lexical information
- The rest of the LRL side is annotated with the lexicalized parser.
  - Several annotations are possible per sentence.
- The annotation that best corresponds with the HRL parse (compared through projection) is retained.
- A new lexicalized LRL parser is trained on the full, disambiguated LRL side of the corpus.

Delexicalization + Annotation projection



- Blue: irrelevant here
- Green: delexicalized parser trained on English
- Yellow: delexicalized parser trained on English, corrected by annotations projected from English

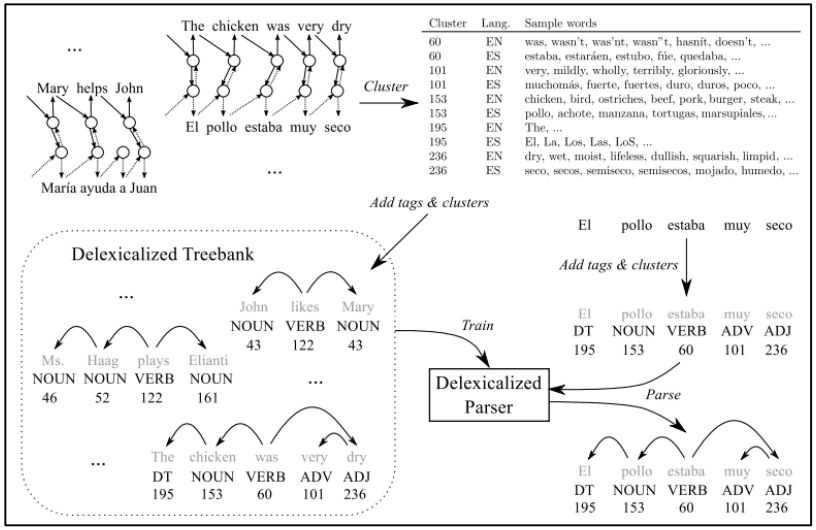
Delexicalization

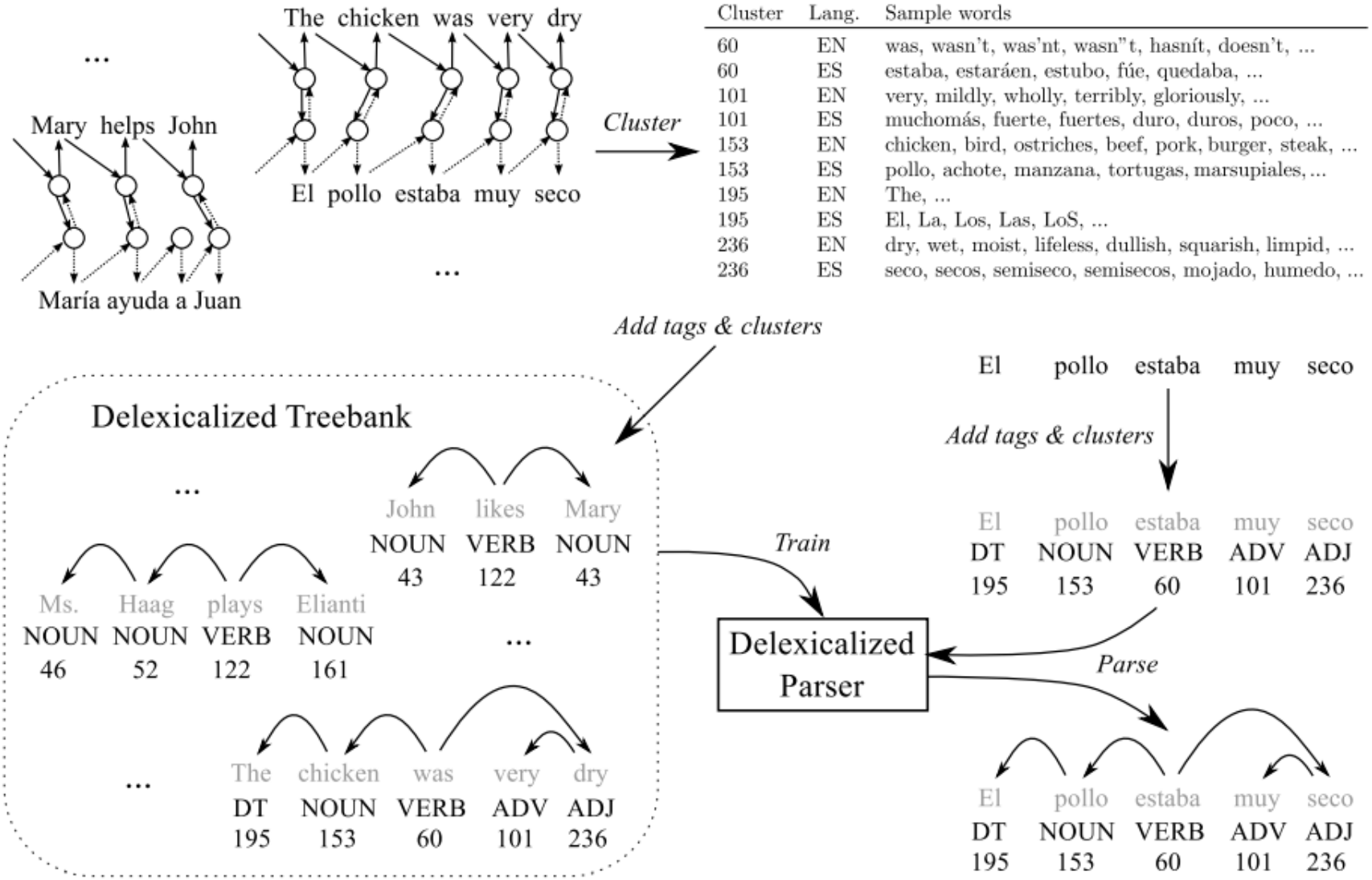
- Delexicalization in itself rarely works well
  - Too much useful information is thrown away
  - Assumptions on language similarity are too strong

- But we can improve delexicalization in several ways:
  - If parallel data are available, we can combine delexicalization with annotation projection
  - We can replace the word forms by abstract word representations that are comparable across languages
  - If we have a bilingual dictionary, we can simply translate the word forms: *relexicalization*

Reading

- Oscar Täckström, Ryan McDonald & Jakob Uszkoreit (2012): *Cross-lingual word clusters for direct transfer of linguistic structure.* Proceedings of NAACL-HLT. http://aclweb.org/anthology/N/N12/N12-1052.pdf
- Question:
  - Try to understand and explain Figure 1.




...
Mary helps John
María ayuda a Juan
The chicken was very dry
El pollo estaba muy seco
...
Cluster
Cluster Lang. Sample words
60 EN was, wasn't, was'nt, wasn"t, hasnít, doesn't, ...
60 ES estaba, estaráen, estubo, fúe, quedaba, ...
101 EN very, mildly, wholly, terribly, gloriously, ...
101 ES muchomás, fuerte, fuertes, duro, duros, poco, ...
153 EN chicken, bird, ostriches, beef, pork, burger, steak, ...
153 ES pollo, achote, manzana, tortugas, marsupiales, ...
195 EN The, ...
195 ES El, La, Los, Las, LoS, ...
236 EN dry, wet, moist, lifeless, dullish, squarish, limpid, ...
236 ES seco, secos, semiseco, semisecos, mojado, humedo, ...
Add tags & clusters
Delexicalized Treebank
John likes Mary
NOUN VERB NOUN
43 122 43
Ms. Haag plays Elianti
NOUN NOUN VERB NOUN
46 52 122 161
The chicken was very dry
DT NOUN VERB ADV ADJ
195 153 60 101 236
Train
Delexicalized Parser
El pollo estaba muy seco
Add tags & clusters
El pollo estaba muy seco
DT NOUN VERB ADV ADJ
195 153 60 101 236
Parse
El pollo estaba muy seco
DT NOUN VERB ADV ADJ
195 153 60 101 236

Cross-lingual word clusters

- Projected clusters
  - Create clusters for language A (256) based on prediction from previous word
  - Assign to each word of language B the cluster of the most often aligned word of language A
    - Word-aligned parallel corpus required
- Cross-lingual clusters
  - Create clusters for language A (as above)
  - Project clusters to language B (as above)
  - Recreate clusters for language B
  - Project clusters to language A

Results

- Source language: English
- Measure: UAS

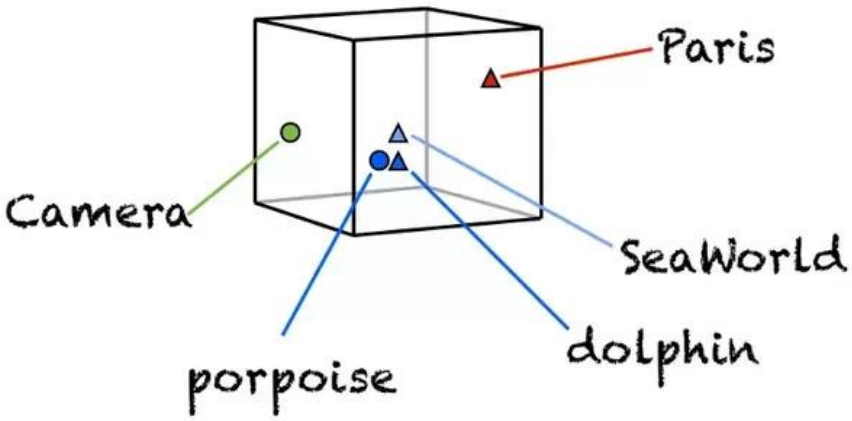
	DA	DE	EL	ES	FR	IT	NL	PT	RU	SV	AVG
NO CLUSTERS	36.7	48.9	59.5	60.2	70.0	64.6	52.8	66.8	29.7	55.4	*54.5*
PROJECTED CLUSTERS	38.9	50.3	61.1	62.6	71.6	68.6	54.5	70.7	32.9	57.0	*56.8*
X-LINGUAL CLUSTERS	38.7	50.7	63.0	62.9	72.1	68.8	54.3	71.0	34.4	56.9	*57.3*

Relexicalization using cross-lingual word representations

- This model still needs large parallel corpora to synchronize the cluster IDs.
  - We might as well project the word forms along the alignments.
- Can we do something similar without (large) parallel corpora?
- Can we use different representations than clusters?
  - For example… word embeddings?

Word embeddings

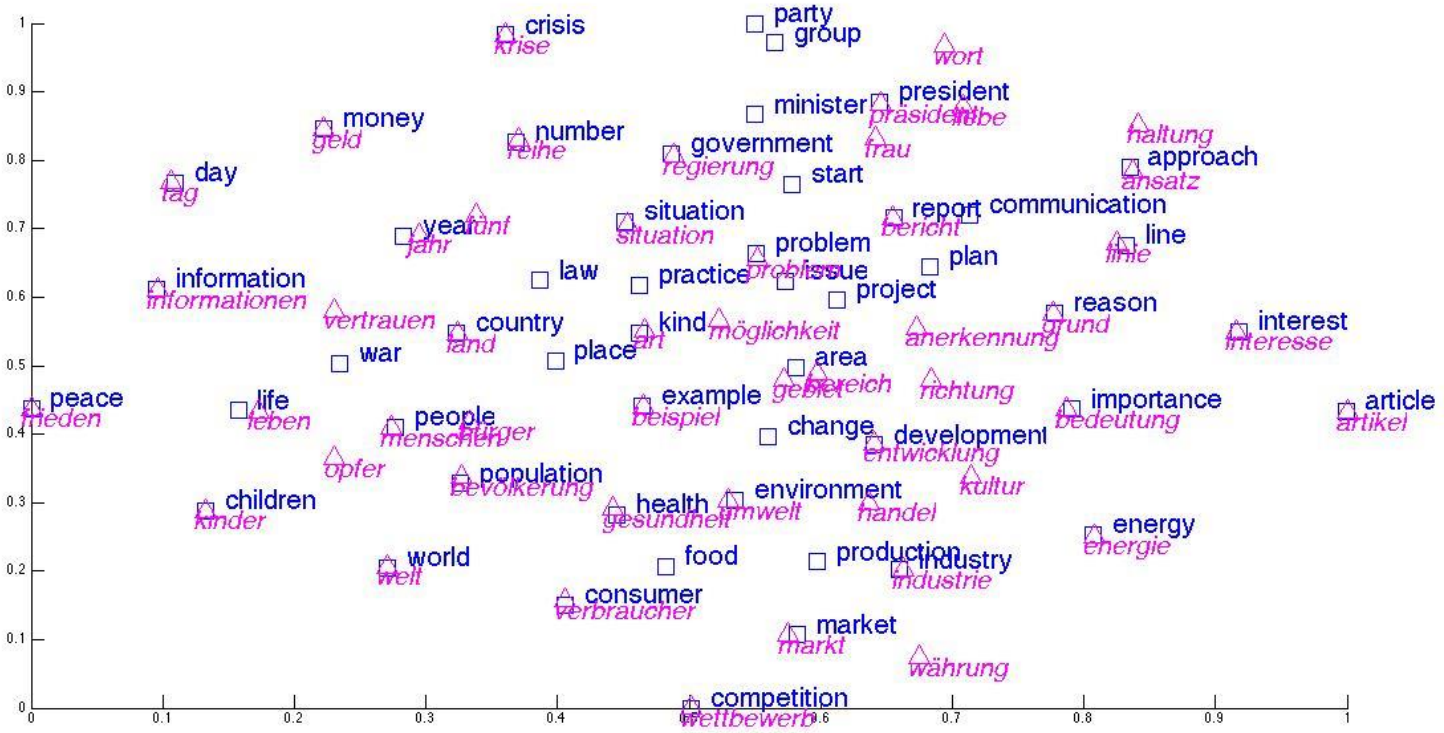
- Represent each word as a multidimensional vector
  - Each word is defined by its position in a multidimensional space
- Words that appear in similar contexts will have similar vectors
  - A large monolingual corpus is sufficient to infer word embeddings



https://www.quora.com/What-does-the-word-embedding-mean-in-the-context-of-Machine-Learning

Bilingual word embeddings

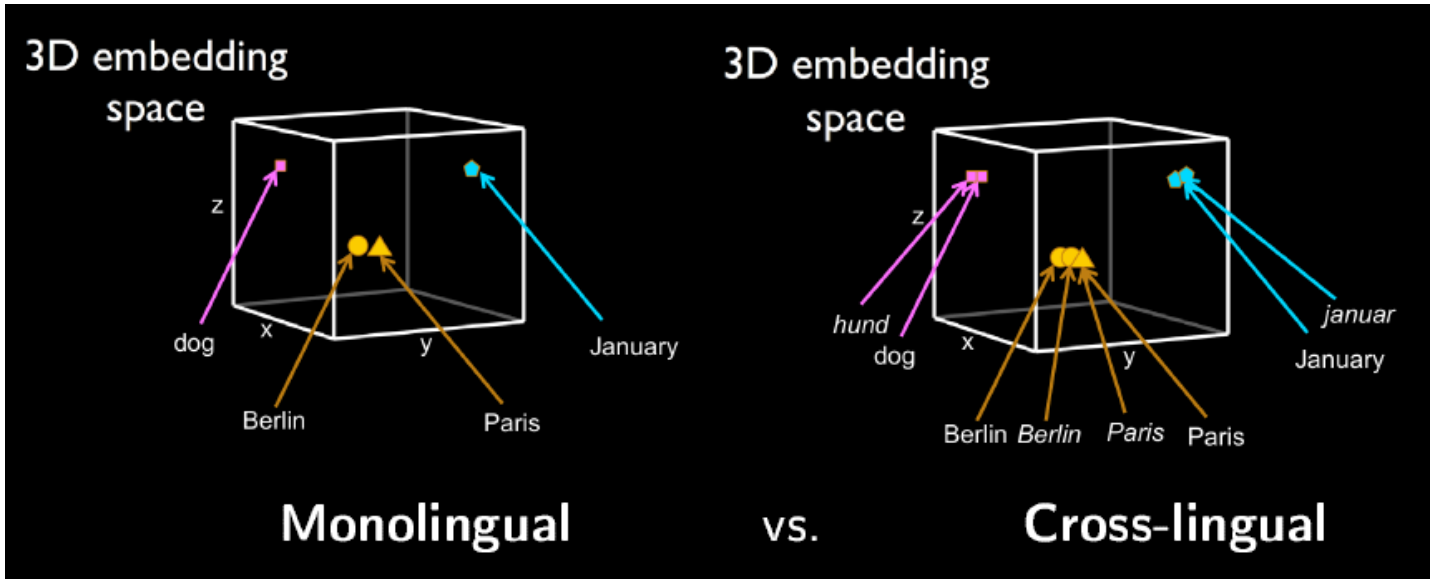
- Words that are translations of each other should be close to each other in the embedding space



http://www.cs.stanford.edu/~lmthang/bivec/biplot.jpg

Bilingual word embeddings

- How do we make sure that the two vector spaces are comparable?



http://people.ds.cam.ac.uk/iv250/tutorial/xlingrep-tutorial.pdf

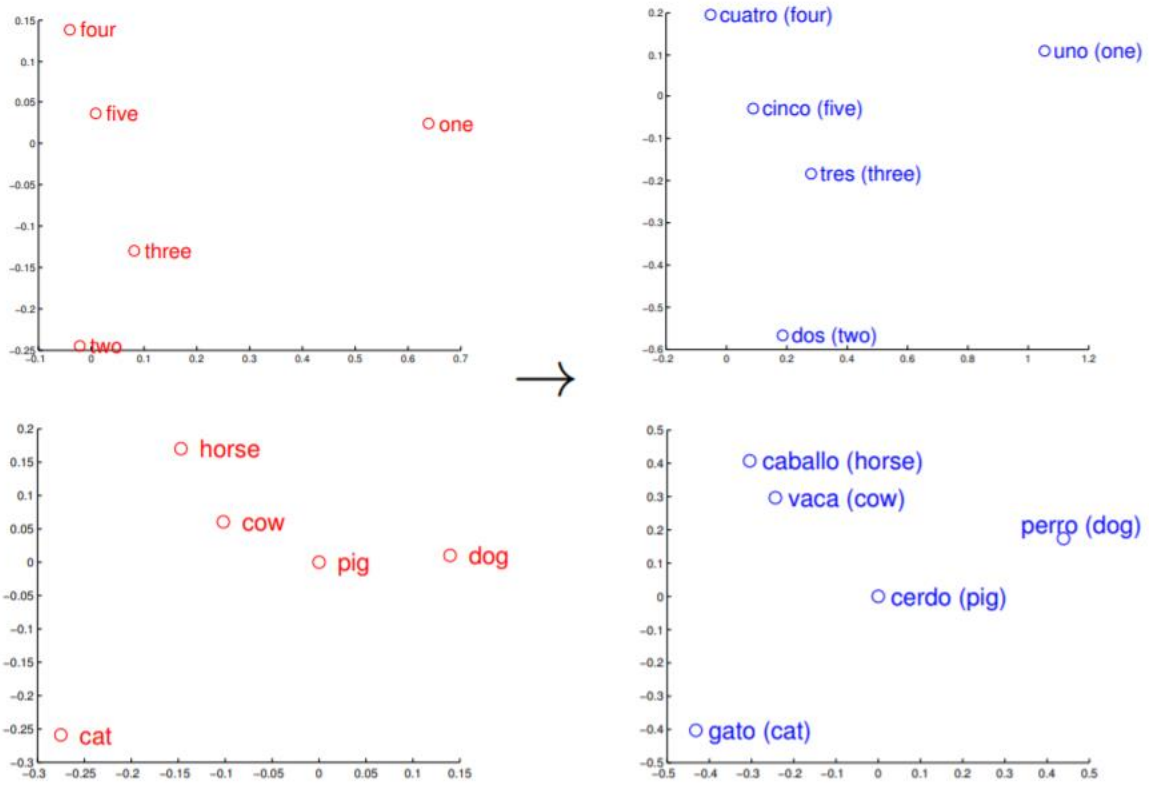
Bilingual word embeddings

- A huge amount of research
  - Cf. EMNLP 2017 tutorial on cross-lingual word representations http://people.ds.cam.ac.uk/iv250/tutorial/xlingrep-tutorial.pdf

- Two main approaches:
  - Learn two independent vector spaces, then learn mapping function
  - Learn single vector space by mixing input data

Bilingual word embeddings Mapping approach

- The geometric structures of vector spaces look similar across languages:

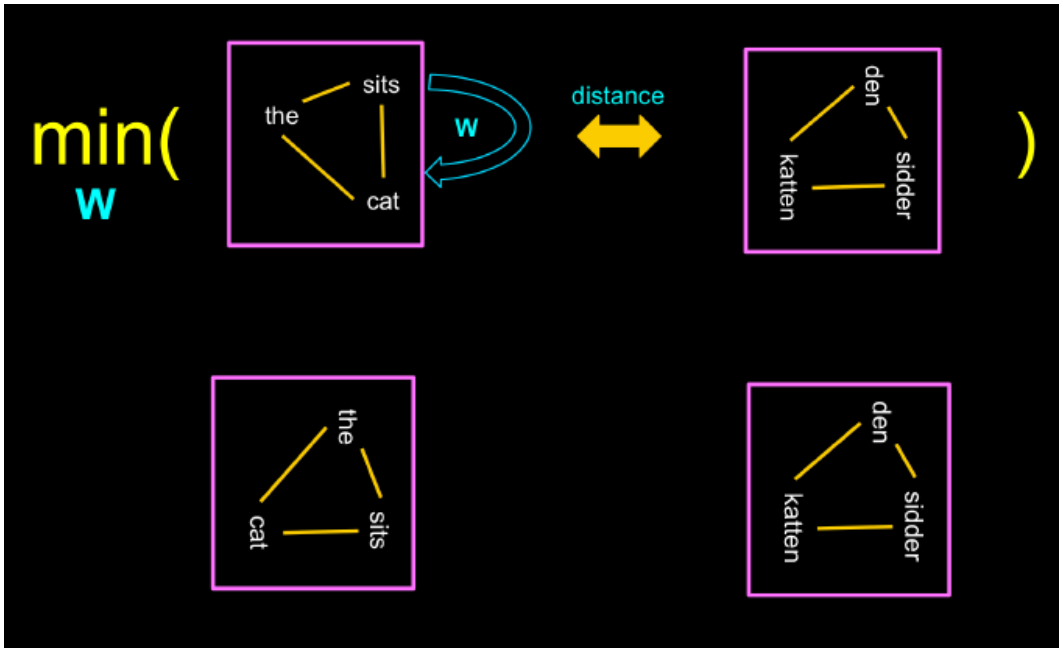


T. Mikolov et al. (2013): *Exploiting similarities among languages for machine translation*. ArXiv.

Bilingual word embeddings Mapping approach

- Learn to transform the pre-trained source language embeddings into a space where the distance between a word and its translation is minimized -> requires translation pairs

T. Mikolov et al. (2013): *Exploiting similarities among languages for machine translation*. ArXiv.



Bilingual word embeddings Data mixing approach

- Gouws & Søgaard (2015): *Simple task-specific bilingual word embeddings*. Proceedings of NAACL-HLT.
  - Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives (BARISTA)
- Algorithm:
  - Replace ½ occurrences of words in A source corpus with B translations
  - Replace ½ occurrences of words in B source corpus with A translations
  - Concatenate both corpora, train word embeddings

build the house
=> build **la** house

construire la maison
=> construire la **house**

Bilingual word embeddings Data mixing approach

- This presupposes that we can translate words from A to B and from B to A
  - We can extract word translations from Wiktionary
  - We don't need exact translations, only *task-specific equivalences*

- Task-specific equivalences
  - POS: *cat* is equivalent to *dog* (animate nouns)
  - NER: *Trump* is equivalent to *Putin* (president names)
- Cross-lingual task-specific equivalences
  - POS: *cat* is equivalent to *Hund*
  - NER: *Trump* is equivalent to *Poutine*

Bilingual word embeddings Data mixing approach

- Task: POS-tagging
  - Delexicalized with embeddings
- POS-specific equivalences
- 50 or 300 dimensions for word embeddings
- DP: Das & Petrov, using bilingual corpora

Language	TC-Perc	Random	Klmtv	POS-50	POS-300	Tr-50	Tr-300	DP
Spanish		81.8	79.8	82.4	81	81.6	82.6	**84.2**
German		82.7	82.8	81.8	84.1	82.6	**84.8**	82.8
Danish		68.9	-	68.9	72.4	71.8	78.4	**83.2**
Swedish	71.6	73.7	-	75	76	75.4	77.5	**80.5**
Italian	80.1	81.3	-	82.1	80.9	82.1	80.7	**86.8**
Dutch	74.5	77.2	-	78.3	77.4	78.7	**80.3**	79.5
Portuguese	76.9	78.1	-	77.3	76.1	80.6	80.5	**87.9**
Avg	75.3	77.7	-	78	78.3	79	80.7	**83.6**

Random embeddings

Bilingual word embeddings Data mixing approach

- Task: POS-tagging
  - Delexicalized with embeddings
- True translational equivalences
- 50 or 300 dimensions for word embeddings
- DP: Das & Petrov, using bilingual corpora

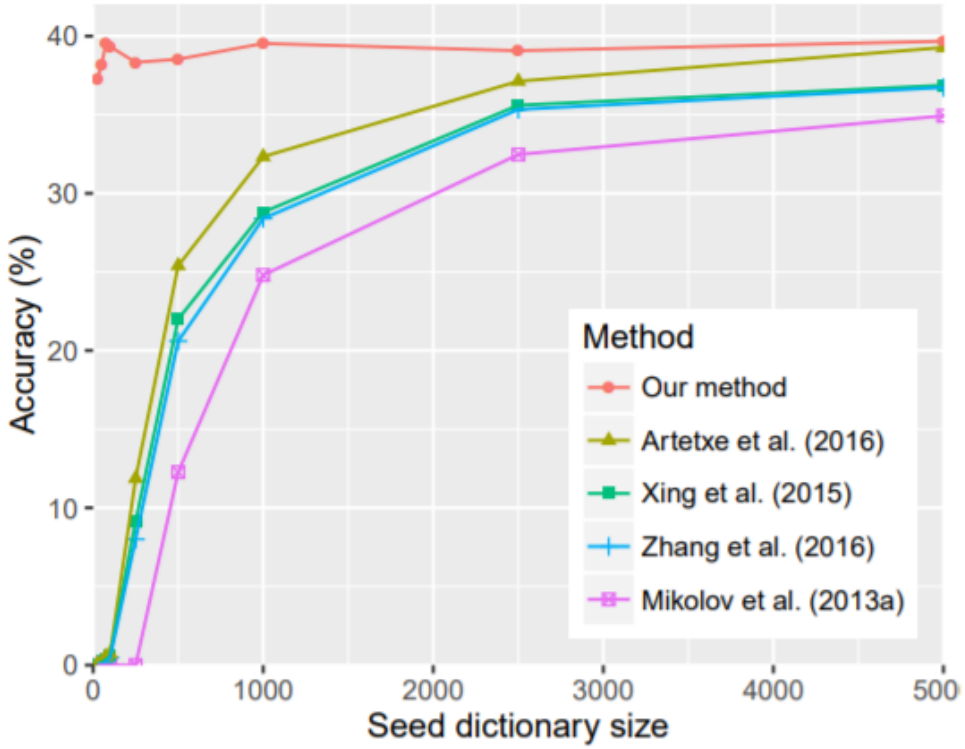
Language	TC-Perc	Random	Klmtv	POS-50	POS-300	Tr-50	Tr-300	DP
Spanish	80.6	81.8	79.8	82.4	81	81.6	82.6	**84.2**
German	80.4	82.7	82.8	81.8	84.1	82.6	**84.8**	82.8
Danish	63	68.9	-	68.9	72.4	71.8	78.4	**83.2**
Swedish	71.6	73.7	-	75	76	75.4	77.5	**80.5**
Italian	80.1	81.3	-	82.1	80.9	82.1	80.7	**86.8**
Dutch	74.5	77.2	-	78.3	77.4	78.7	**80.3**	79.5
Portuguese	76.9	78.1	-	77.3	76.1	80.6	80.5	**87.9**
Avg	75.3	77.7	-	78	78.3	79	80.7	**83.6**

Bilingual word embeddings

- What can we do if bilingual dictionaries are not available?
  - M. Artetxe et al. (2017): Learning bilingual word embeddings with (almost) no bilingual data. ACL 2017.
- Go back to the mapping approach
- A small translation lexicon is sufficient
  - 25 word pairs
  - Numeral pairs (trivially generated)
- Use self-learning to gradually increase the lexicon

Bilingual word embeddings

- Results on bilingual lexicon induction task:



M. Artetxe et al. (2017): Learning bilingual word embeddings with (almost) no bilingual data. ACL 2017.

Bilingual word embeddings
Other approaches

	Parallel	Comparable
Word	Mikolov et al. (2013)	Kiela et al. (2015)
	Faruqui & Dyer (2014)	Vulić et al. (2016)
	Xing et al. (2015)	Vulić et al. (2017)
	Dinu et al. (2015)	Zhang et al. (2017)
	Lazaridou et al. (2015)	Hauer et al. (2017)
	Zhang et al. (2016)	
	Zou et al. (2013)	
	Ammar et al. (2016)	
	Artexte et al. (2016)	
	Xiao & Guo (2014)	
	Gouws and Søgaard (2015)	
	Duong et al. (2016)	
	Gardner et al. (2015)	
	Smith et al. (2017)	
	Mrkšić et al. (2017)	
	Artetxe et al. (2017)	

Delexicalization

- Delexicalization in itself rarely works well
  - Too much useful information is thrown away
  - Assumptions on language similarity are too strong

- But we can improve delexicalization in several ways:
  - If parallel data are available, we can combine delexicalization with annotation projection
  - We can replace the word forms by abstract word representations that are comparable across languages
  - If we have a bilingual dictionary, we can simply translate the word forms within the trained model: *relexicalization*

Relexicalization

- Train a tagger on annotated HRL data
  - Tag-word associations → Translate words to LRL
  - Tag sequences → Keep unchanged

ecologista	AQ 5	
ecologistas	AQ 4	NC 1
ecología	NC 3	
ecológica	AQ 3	
ecológicas	AQ 1	
ecológico	AQ 8	
ecológicos	AQ 2	

DT NC	156
DT AQ	23
AQ NC	34
…	

ecolochía	NC 3	
ecolochica	AQ 3	
ecolochicas	AQ 1	
ecolochico	AQ 8	
ecolochicos	AQ 2	
ecolochista	AQ 5	
ecolochistas	AQ 4	NC 1

DT NC	156
DT AQ	23
AQ NC	34
…	

Relexicalization

- In order to translate the words in the model parameter file, we need a bilingual dictionary
  - Extracted from Wikipedia or Wiktionary
  - Extracted by word alignment from a parallel corpus (Zeman & Resnik 2008)
  - Generated by matching cognates

    A. Feldman, J. Hana & C. Brew (2006): *A cross-language approach to rapid creation of new morphosyntactically annotated resources.* LREC 2006.
- This is a whole area of research: *Bilingual lexicon induction*

Relexicalization using morphological analysis

- Train a tagger on annotated HRL data
  - Tag-word associations → Replace with LRL words
  - Tag sequences → Keep unchanged

ecologista	AQ 5	
ecologistas	AQ 4	NC 1
ecología	NC 3	
ecológica	AQ 3	
ecológicas	AQ 1	
ecológico	AQ 8	
ecológicos	AQ 2	

tamién	RG 257	RN 257
sieglu	NC 35	
ecolochía	NC 23	VS 23
dende	RG 18	SN 18
primer	AQ 14	AO 14
Asturies	NP 46	
asturianu	AQ 79	

DT NC	156
DT	
AQ	
...	

Tag associations guessed with a «morphological analyzer»

DT NC	156
...	

Word list + frequencies extracted from monolingual corpus

Relexicalization using morphological analysis

- Train a tagger on annotated HRL data
  - Tag-word associations → Replace with LRL words
  - Tag sequences → Keep unchanged

ecologista	AQ 5	
ecologistas	AQ 4	NC 1
ecología	NC 3	
ecológica	AQ 3	
ecológicas	AQ 1	
ecológico	AQ 8	
ecológicos	AQ 2	

DT NC	156
DT AQ	23
AQ NC	34
...	

tamién	RG 257	RN 257
sieglu	NC 35	
ecolochía	NC 23	VS 23
dende	RG 18	SN 18
primer	AQ 14	AO 14
Asturies	NP 46	
asturianu	AQ 79	

DT NC	156
...	

Uniform frequencies for ambiguous words

Relexicalization using cognate counts

- Train a tagger on annotated HRL data
  - Tag-word associations → Replace with LRL words
  - Tag sequences → Keep unchanged

ecologista	AQ 5	
ecologistas	AQ 4	NC 1
ecología	NC 3	
ecológica	AQ 3	
ecológicas	AQ 1	
ecológico	AQ 8	
ecológicos	AQ 2	

tamién	RG 257	RN 257
sieglu	NC 35	
ecolochía	NC 35	VS 12
dende	RG 18	SN 18
primer	AQ 14	AO 14
Asturies	NP 46	
asturian	AQ 79	

DT NC	156
DT AQ	23
AQ NC	34
...	

Formula for computing counts:

$$p'_r(t) = \frac{p_s(t) + p_r(t)}{2}$$

Feldman et al.

- Plain model transfer as baseline
  - Spanish tagger applied to Catalan: 36.5%
  - Spanish tagger applied to Portuguese: 56.9%
  - Czech tagger applied to Russian: 45.6%
- Relexicalization using morphological analyzer:
  - Spanish-Catalan: 70.7%
  - Spanish-Portuguese: 77.2%
  - Czech-Russian: 78.6%
- Relexicalization using cognates:
  - Spanish-Catalan: 75.2%
  - Spanish-Portuguese: 82.1%
  - Czech-Russian: 80.4%

Model transfer

- Assumptions regarding the linguistic proximity between HRL and LRL?
  - Word order
  - Polysemy
  - Homonymy

Tomorrow...

Readings

- Yves Scherrer & Achim Rabus (2017): *Multi-source morphosyntactic tagging for Spoken Rusyn*. Proceedings of VarDial 2017.
  - Plain model transfer using multiple high-resource languages
- Delphine Bernhard & Anne-Laure Ligozat (2013): *Hassle-free POS-Tagging for the Alsatian Dialects*. In: Marcos Zampieri & Sascha Diwersy: Non-Standard Data Sources in Corpus Based-Research, Shaker, ZSM Studien.
  - “Minimalistic relexicalization by hand”