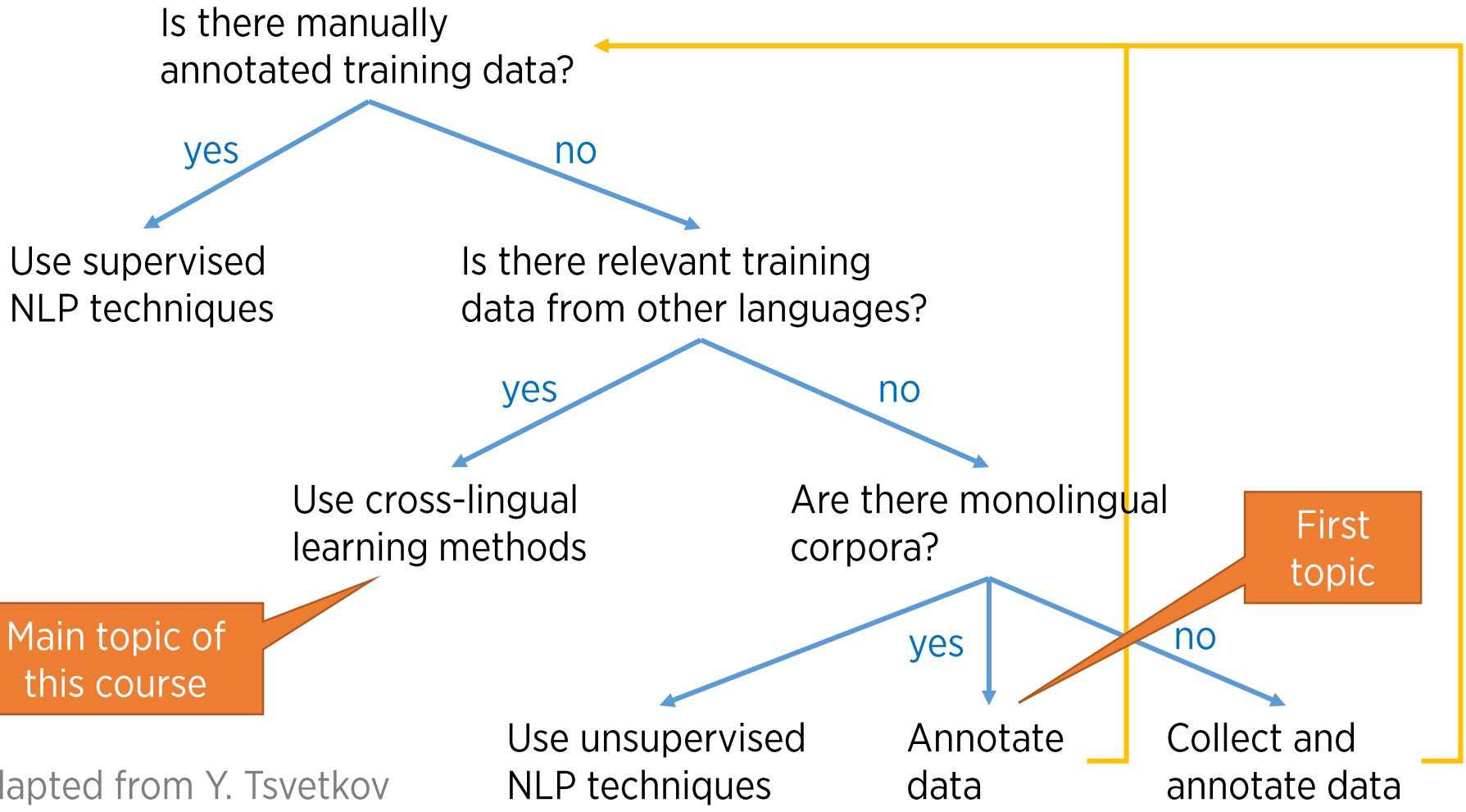
Language technology for low-resource languages

Day 2/5

LOT 2018, Groningen

Yves Scherrer, University of Helsinki

What resources can we get?



Adapted from Y. Tsvetkov

If you don’t have resources…

… create some!

Case study: Learning taggers from 2 hours of annotation

- How good a tagger can one get with only 2 hours of annotation work?
  - Assumption: monolingual raw data is available
- Tagging is a sequence labeling task:
  - Every item (word) of a sequence (sentence) gets a label (part-of-speech)
- What is the best data annotation strategy?
  - *Token supervision:* Annotate complete sentences
  - *Type supervision:* Annotate the most frequent words in isolation
- Can the results be improved with appropriate machine learning methods?

Case study: Learning taggers from 2 hours of annotation

- A (simple) tagger needs to learn two things:
  - Emissions: the most probable associations between words and tags
  - Transitions: the most probable sequences of tags
- Approach 1: annotate a corpus ("tokens")
  - Emissions can be learned easily, and frequency information is available
  - Downside: the same words end up being annotated several times
  - Transitions can be learned easily
- Approach 2: annotate a word list ("types")
  - Emissions can be learned easily, but frequency information is not readily available (in case of ambiguity)
  - More efficient, as each form is annotated only once
  - Transitions cannot be learned easily (needs 2 passes over data)

Experimental setup

	Linguist A		Linguist B	
	Kinyarwanda (Rwanda)	**English**	**Malagasy (Madagascar)**	**English**
2h token annotation	90 sentences 1537 tokens 750 types	86 sentences 1897 tokens 903 types	92 sentences 1805 tokens 666 types	107 sentences 2650 tokens 959 types
2h type annotation	1798 types	1644 types	1067 types	1090 types

- Kinyarwanda 14 tags, Malagasy 24 tags, English 45 tags
- Additional unannotated data: 100 000 words/language
- Annotated sentences for evaluation
  (by the same linguists, but not included in the 2h-tasks)

Experiment 0

- Train a tagger with the ~100 annotated sentences:

Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

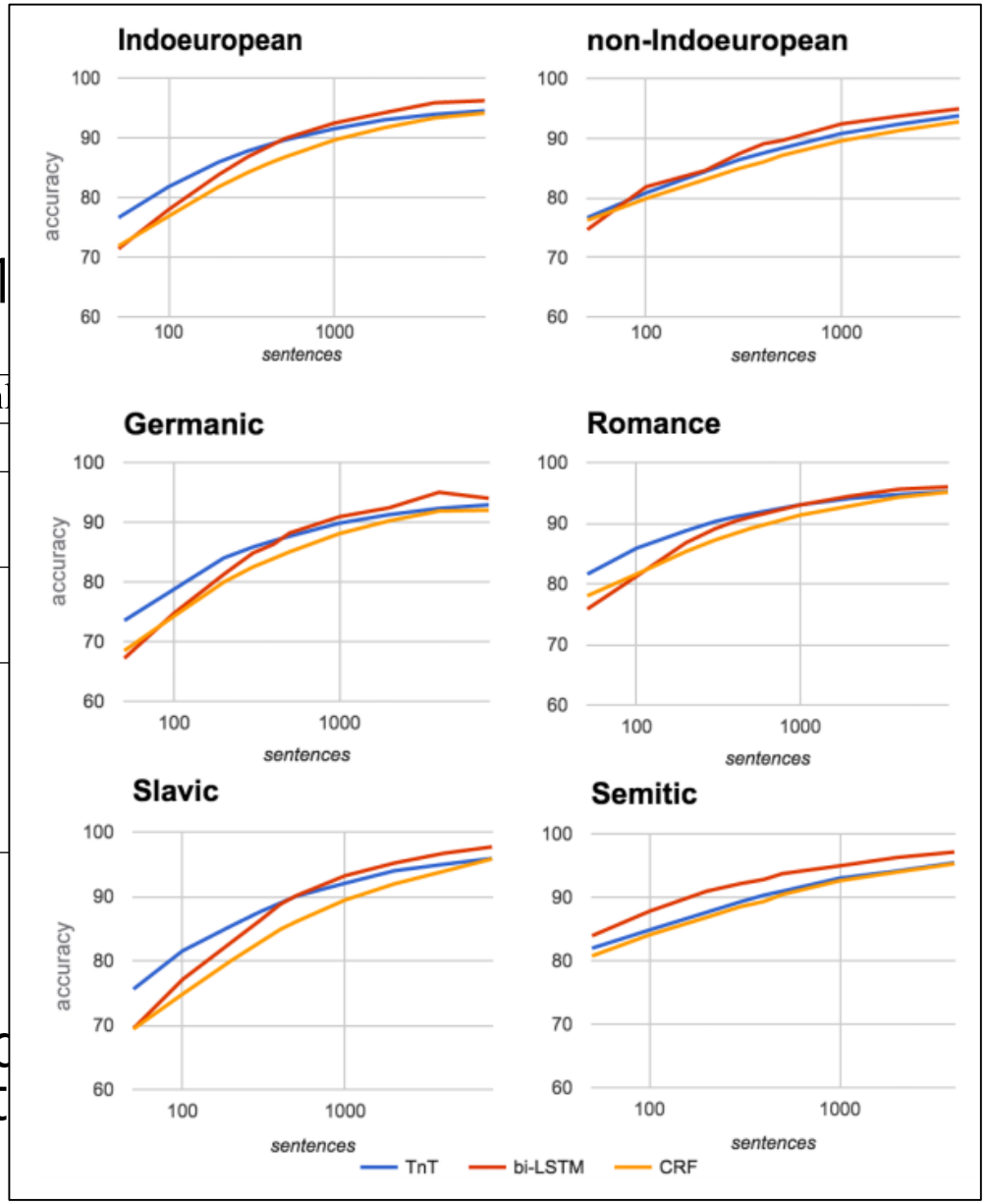
- T = Total
  K = Known words (seen during training)
  U = Unknown words (not seen during training)

Experiment 0

- Train a tagger with the ~1

Human Annotations	0. No EM			1. EM on	
Initial data	T	K	U	T	K
KIN tokens A	**72**	90	58	55	82
KIN types A				63	77
MLG tokens B	74	89	49	68	87
MLG types B				71	87
ENG tokens A	63	83	38	62	83
ENG types A				66	76
ENG tokens B	70	87	44	70	87
ENG types B				69	83

- T = Total
  K = Known words (seen c
  U = Unknown words (not



Experiment 1

- Train tagger on word list and unannotated data
  - Constrain prediction for known words (those in the word list)
  - Assume uniform tag distribution for unknown words
  - Several passes over the text using EM (Expectation Maximisation)

Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

Experiment 1

- Train a tagger with the 100 annotated sentences and the unannotated sentences
  - Annotate the raw text with the 100-sentence tagger
  - Optimize tag associations and emissions in several passes through the raw text using EM

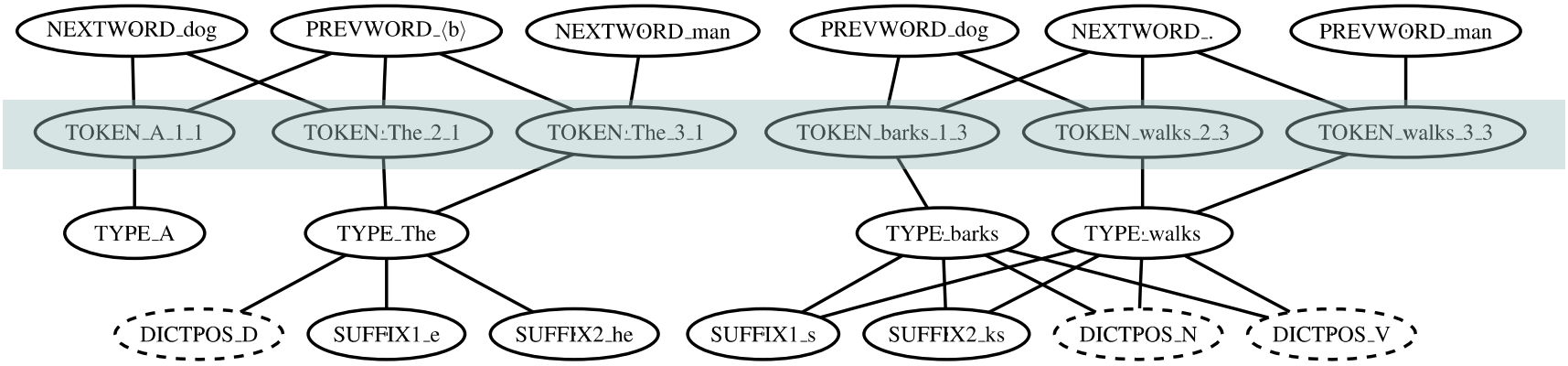
Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

Tag dictionary expansion: Label propagation

- For a lot of words, the possible tags are unknown.
- Adding unannotated data does not help much in this respect. (Why?)
- How can we determine the most probable tags for unknown words?
  - Label propagation: a machine learning algorithm to transfer labels (=tags) to similar words
  - What are similar words?
    - Words with the same suffix or prefix
    - Words with the same predecessor or successor in a sentence
  - All these properties are inserted into a graph.
    - Tags are assigned to the known words and then transferred onto the similar words.

Tag dictionary expansion: Label propagation

- A dog barks.
- The dog walks.
- The man walks.



Experiment 2

- LP computes a tag distribution for every word
- The unannotated corpus is annotated with the tags found by LP, then a new tagger is trained.

Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

Experiment 2

- LP computes a tag distribution for every word
- The unannotated corpus is annotated with the tags found by LP, then a new tagger is trained.

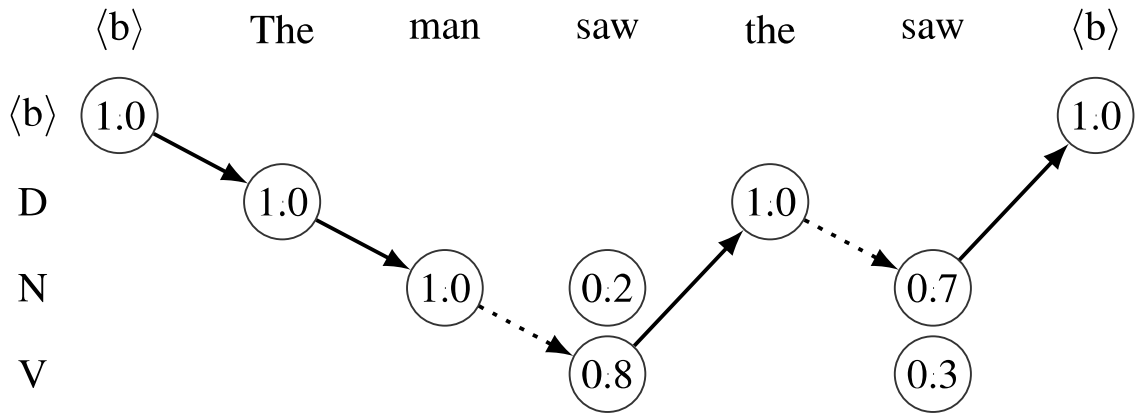
Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

Model minimization

- Label propagation introduces a lot of noise (i.e. wrong tags)

- A lot of tag combinations (of two adjacent words in the text) never occur.
  - Goal: find the minimal set of tag combinations
  - Tags are weighted according to the LP result
  - The tag combinations with the highest weight are added

Model minimization



Experiment 3

- Model minimization creates an optimal tag sequence for each unannotated sentence.
- A new tagger is trained on this data.

Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

Experiment 3

- Model minimization creates an optimal tag sequence for each unannotated sentence.
- A new tagger is trained on this data.

Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

Examples

for	*IN	*RP	JJ	NN	CD
(1) EM	1,221	2764		9	5
(2) LP	4,003				
(3) min	4,004		1		
gold	3,999	5			

, (comma)	*,	*:	JJS	PTD	VBP
(1) EM	24,708		4	3	3
(2) LP	15,505	9226			1
(3) min	24,730				
gold	24,732				

opposition	NN	JJ	DT	NNS	VBP
(1) EM	24	4	1	4	4
(2) LP	41	4			
(3) min	45				
gold	45				

External dictionary

- Kinyarwanda:
  - 3700 entries (kinyarwanda.net)
- Malagasy:
  - 78′000 entries (malagasyworld.org)
- English:
  - 614′000 entries (Wiktionary)

- Wiktionary is too small for Kinyarwanda (9 entries) and for Malagasy (3365 entries).
- Dictionary data is incorporated into the LP graph.

Experiment 4

- Include the external dictionary in the label propagation graph, train new tagger on this data
- Conclusion: spending more time on enlarging the dictionary may not be helpful…

Human Annotations	0. No EM			1. EM only			2. With LP			3. LP+min			4. LP(ed)+min		
Initial data	T	K	U	T	K	U	T	K	U	T	K	U	T	K	U
KIN tokens A	**72**	90	58	55	82	32	71	86	58	71	86	58	71	86	58
KIN types A				63	77	32	78	83	69	**79**	83	70	**79**	83	70
MLG tokens B	74	89	49	68	87	39	74	89	49	74	89	49	**76**	90	53
MLG types B				71	87	46	72	81	57	74	86	56	**76**	86	60
ENG tokens A	63	83	38	62	83	37	**72**	85	55	**72**	85	55	**72**	85	56
ENG types A				66	76	37	75	81	56	**76**	83	56	74	81	55
ENG tokens B	70	87	44	70	87	43	**78**	90	60	**78**	90	60	**78**	89	61
ENG types B				69	83	38	75	82	61	**78**	85	61	**78**	86	61

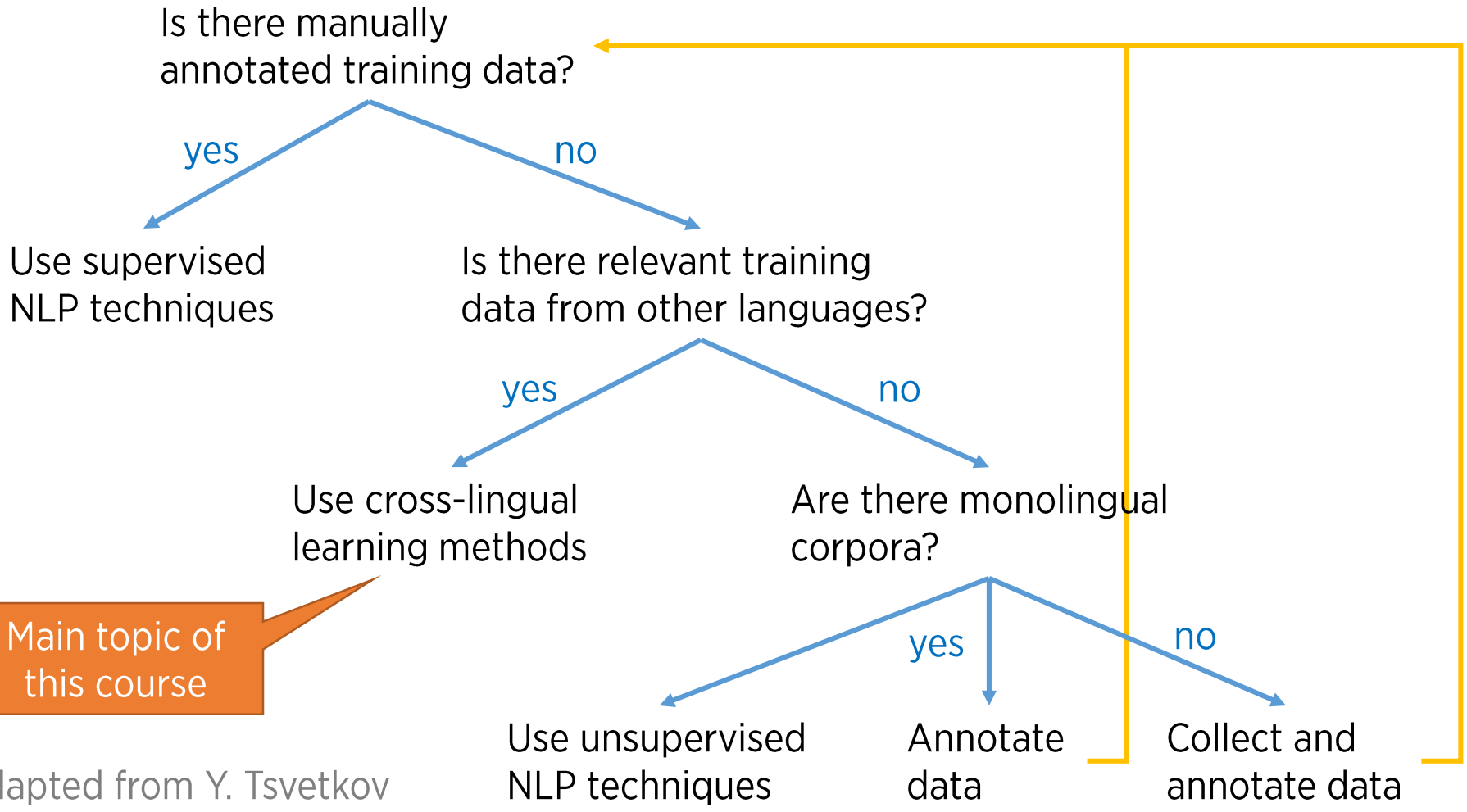
Conclusions

- What is more efficient, type or token annotations?
  - Annotators have personal preferences
  - In general, it is more efficient to create type annotations (word lists)
    - No repetitions, thus better coverage in the same period
    - Focus on most frequent (thus most useful) words
    - Tag frequencies and distributions can also be estimated from an unannotated corpus (and the appropriate machine learning techniques)

Cross-lingual learning

Transfer learning

What resources can we get?



Adapted from Y. Tsvetkov

Cross-lingual learning / Transfer learning

- We have the data to train a model for language A, but need a model for language B
  - Languages A and B may be related, but not necessarily
  - Different methods, depending on where in the process the switch/transfer from A to B occurs

- A = high-resource language (HRL)
- B = low-resource language (LRL)

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

Today

Tomorrow

Cross-lingual learning techniques

Example task:
Part-of-speech tagging

- Data transfer – the ideal setup:
  - Word-aligned parallel corpus whose A side is manually annotated:

ADP	DET	NOUN	AUX	DET	NOUN	PUNCT
In	the	beginning	was	the	word	.
Am		Anfang	war	das	Wort	.

- Task:
  - Copy all annotations along alignment links to B
    - Easier said than done…
  - Train model on annotated B side

Cross-lingual learning techniques

- Data transfer – the common setup:
  - Annotated monolingual corpus of language A:

ADV	DET	NOUN	AUX	ADV	ADJ	PUNCT
Maybe	the	dresscode	was	too	stuffy	.

  - Word-aligned but unannotated parallel corpus:

In	the	beginning	was	the	word	.
Am		Anfang	war	das	Wort	.

- How can we connect the two?

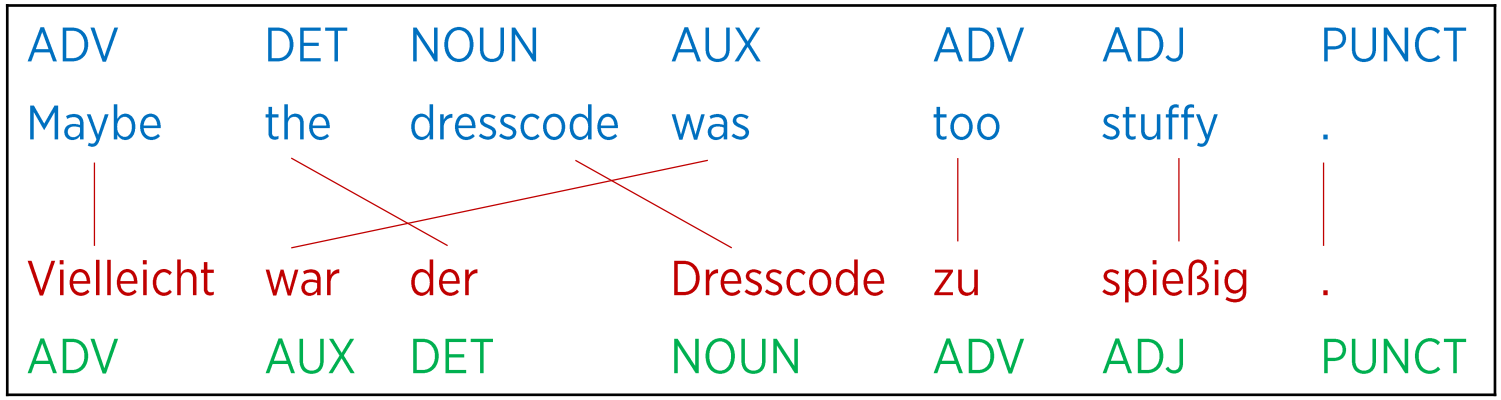
Cross-lingual learning techniques

- Data transfer – annotation projection:
  - Parallel data set and annotated A corpus are disjoint
  - Train A model on annotated A corpus
  - Annotate A side of parallel data with it
  - Copy all annotations along alignment links to B
  - Train model on annotated B side

ADP	DET	NOUN	AUX	DET	NOUN	PUNCT
In	the	beginning	was	the	word	.
Am		Anfang	war	das	Wort	.
ADP		NOUN	AUX	DET	NOUN	PUNCT

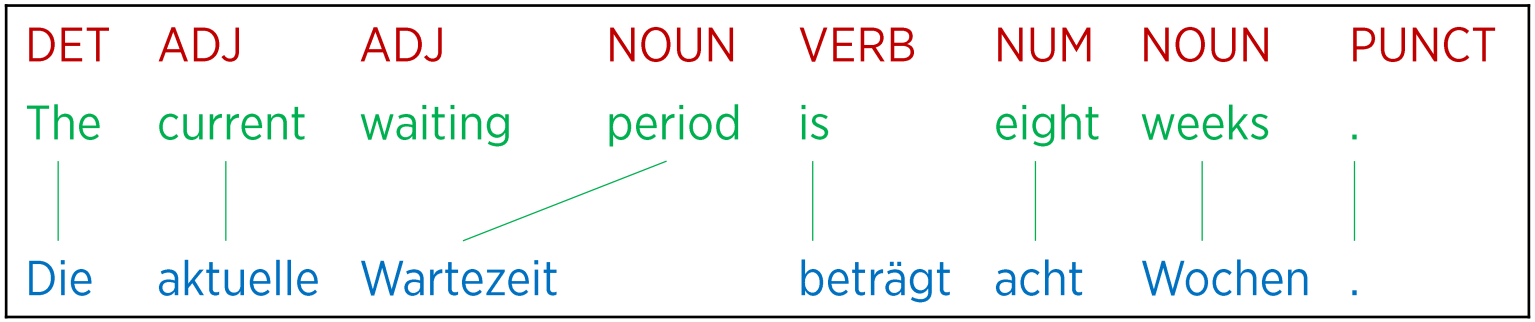
Cross-lingual learning techniques

- Data transfer – training data translation:
  - Parallel data set and annotated A corpus are disjoint
  - Train A-B machine translation model on parallel data
  - Translate annotated A corpus
  - Copy all annotations along alignment links to B
  - Train model on annotated B side



Cross-lingual learning techniques

- Data transfer – test data backtranslation:
  - Parallel data set and annotated A corpus are disjoint
  - Train A model on annotated A corpus
  - Train B-A machine translation model on parallel data
  - Translate B test corpus to A
  - Annotate translated corpus with A model
  - Project annotations back to B

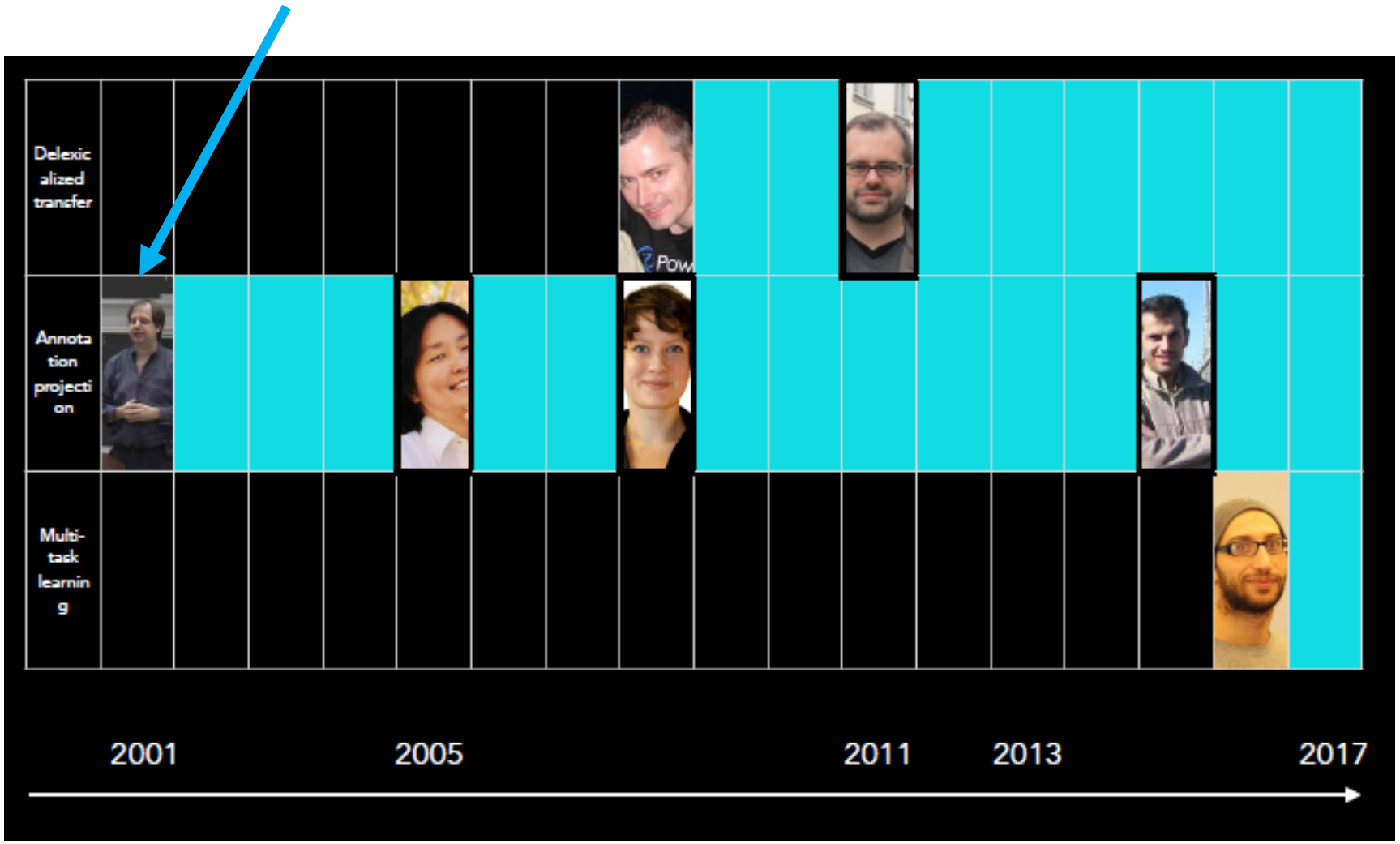


Cross-lingual learning techniques

- Data transfer – test data backtranslation:
  - Why would one choose this option?

- This approach is typical for historical language varieties:
  - A = present-day variety
  - B = historical variety
  - B-A translation = normalization / modernization
- Why?
  - The historical variety is not standardized
  - Translation from B to A is easier than from A to B

Annotation projection



Vulić, Søgaard & Faruqui: Tutorial on cross-lingual word representations, EMNLP 2017.
http://people.ds.cam.ac.uk/iv250/tutorial/xlingrep-tutorial.pdf

Reading

- Pioneering paper on annotation projection:
  - David Yarowsky & Grace Ngai: *Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora*. NAACL 2001.
- Questions:
  - What is the task of word alignment about?
  - What can go wrong when using direct projection, as illustrated by the examples in Figure 2?
  - Annotation projection assumes that it is easier to obtain a parallel corpus (and a tagger for the high-resource language) than a directly annotated corpus for the low-resource language. Do you agree with this assumption?

Word alignment tools

- GIZA++ and Moses helper scripts
  - https://github.com/moses-smt/mgiza
  - http://www.statmt.org/moses/?n=Development.GetStarted
- fast_align
  - https://github.com/clab/fast_align
- efmaral
  - https://github.com/robertostling/efmaral
- Anymalign
  - https://anymalign.limsi.fr/

Word alignment

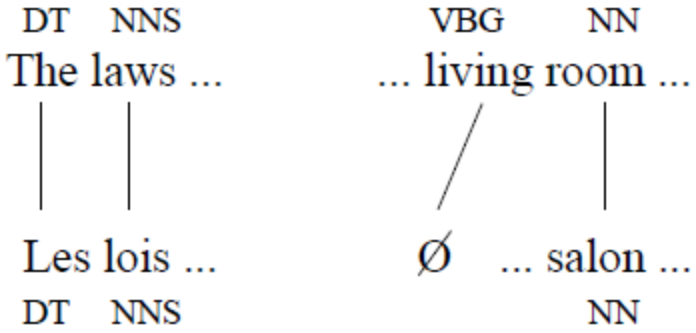
- Example input for *fast_align*:

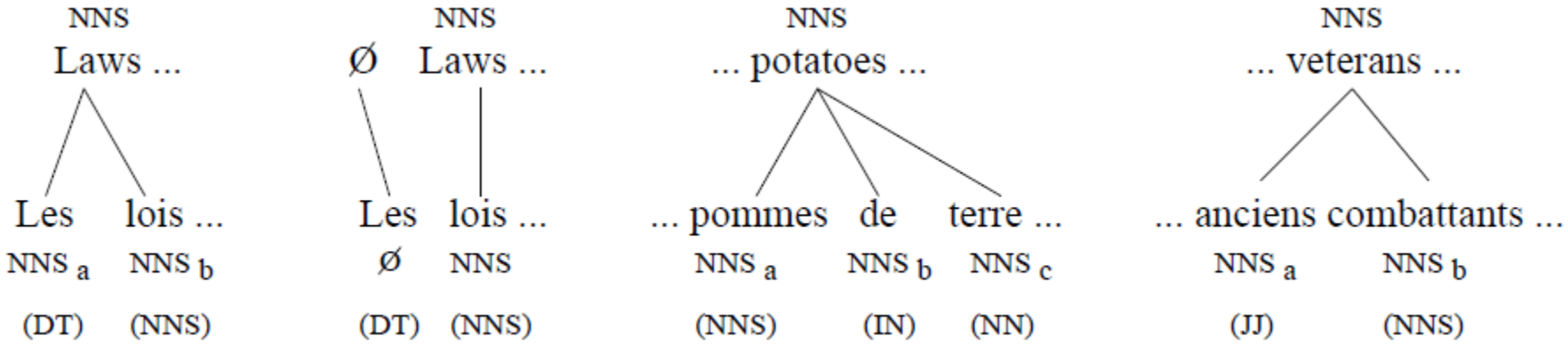
```
doch jetzt ist der Held gefallen . ||| but now the hero has fallen .
neue Modelle werden erprobt . ||| new models are being tested .
doch fehlen uns neue Ressourcen . ||| but we lack new resources .
```

- Example output from *fast_align*:

```
0-0 1-1 2-4 3-2 4-3 5-5 6-6
0-0 1-1 2-2 2-3 3-4 4-5
0-0 1-2 2-1 3-3 4-4 5-5
```

Direct projection





Noise reduction

$$\hat{P}(t_{(2)}|w) = \lambda_1 P(t_{(2)}|w) \quad \text{where } \lambda_1 < 1.0$$
$$\hat{P}(t_{(1)}|w) = 1 - \hat{P}(t_{(2)}|w)$$
$$\hat{P}(t_{(c)}|w) = 0 \quad \text{for all } c > 2$$

Noise reduction techniques are directly applied to the B tagger training process. The projected corpus remains noisy.

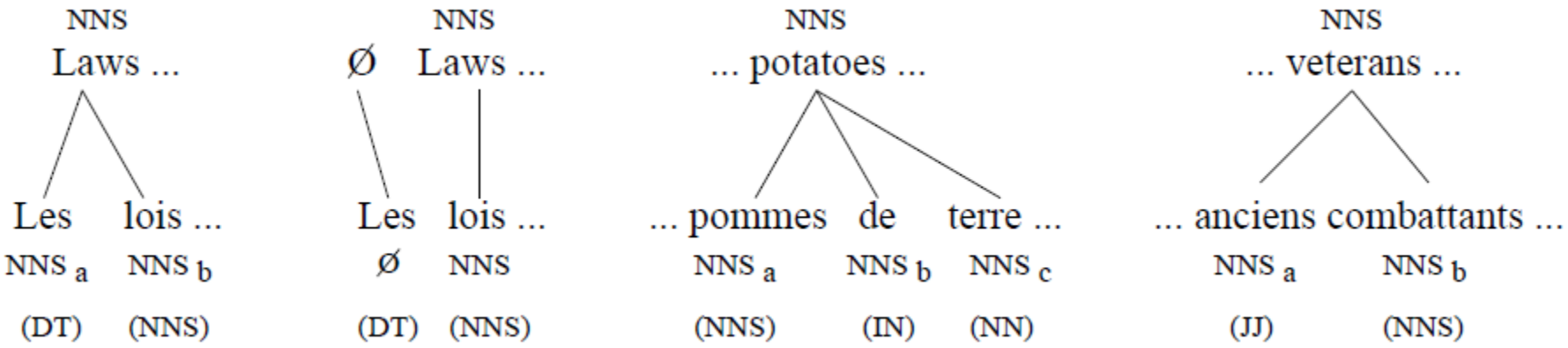
Word	Directly Projected Tag					Tag Error
	J	N	V	R	I	
achat	0	**62**	48	0	1	0.44
cadre	2	**35**	7	1	1	0.27
cadres	1	**5**	0	0	0	0.17
prévu	1	11	**48**	0	0	0.20

Table 1: Raw projected tag distributions.

Word	Smoothed $\hat{P}(t\|w)$					
	N	V	NN	NNS	VBN	VBG
achat	.76	.24	.73	.03	.03	.21
cadre	.90	.10	.86	.04	.03	.00
cadres	.94	.00	.04	.90	.00	.00
prévu	.09	.91	.08	.01	.86	.00

Table 2: Smoothed $\hat{P}(t|w)$ tag probabilities

Noise reduction



$$P(t|w) = \lambda_2 P_{\text{1-to-1}}(t|w) + (1 - \lambda_2) P_{\text{1-to-n}}(t|w)$$

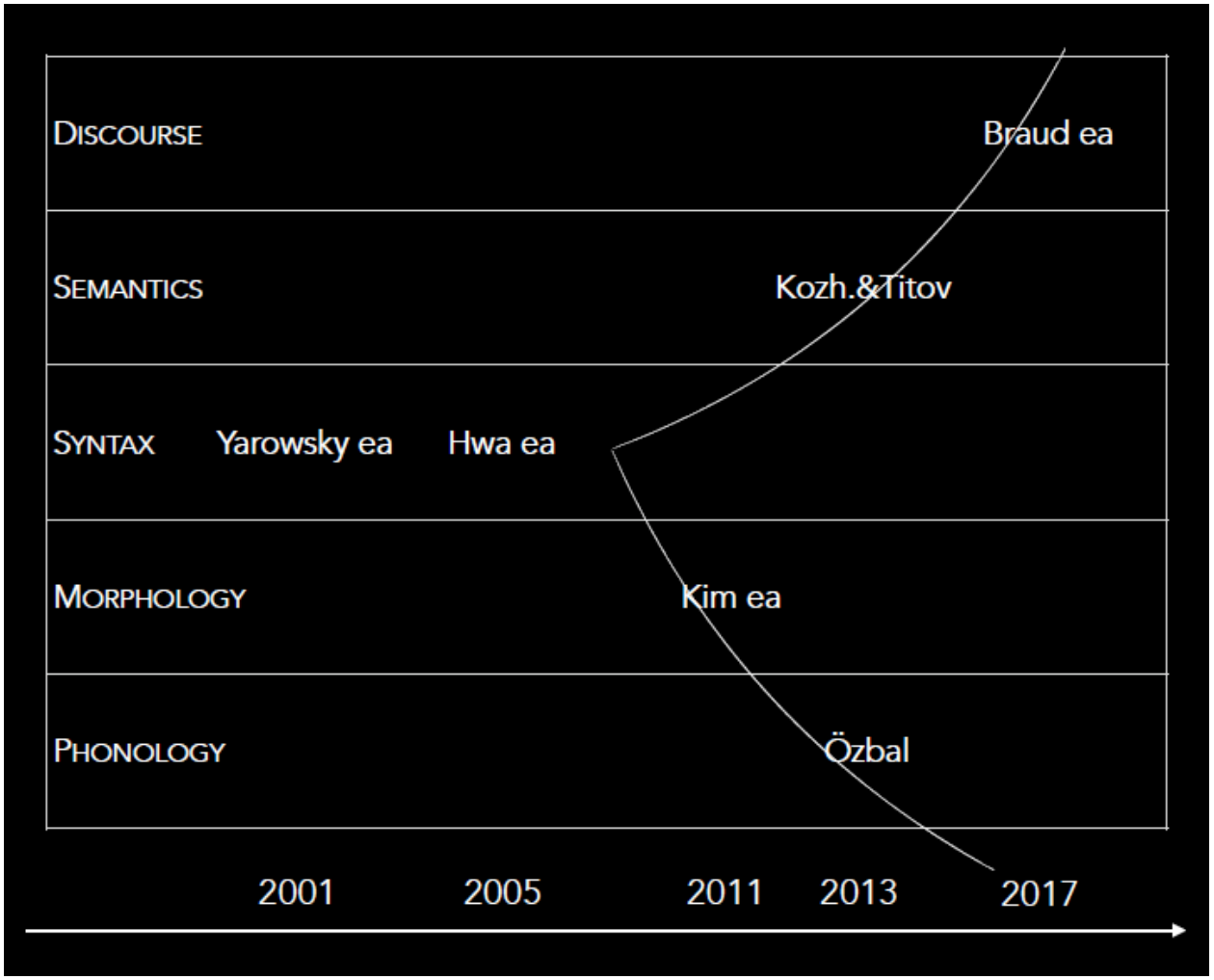
Annotation projection

- Annotation projection assumes that it is easier to obtain a parallel corpus (and a tagger for the high-resource language) than a directly annotated corpus for the low-resource language.
- Do you agree with this assumption?

Annotation projection

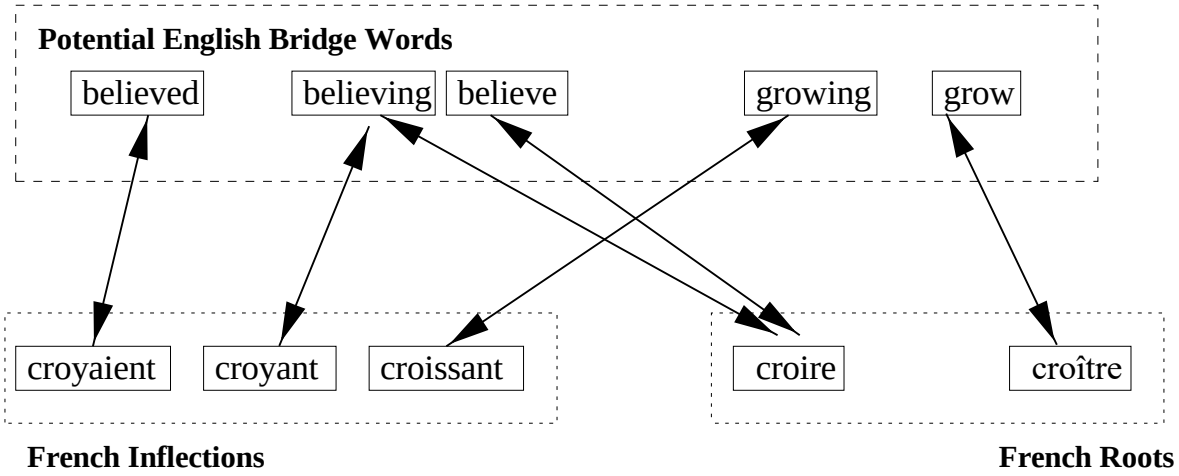
Authors	Task	Languages
Yarowsky & Ngai 2001	POS tagging NP chunking	EN→FR EN→FR, EN→ZH
Yarowsky et al. 2001	Morphology	EN→FR
Riloff et al. 2002	Information extraction	EN→FR
Diab & Resnik 2002	Word sense disambiguation	EN↔FR
Hwa et al. 2005	Dependency parsing	EN→EU (Basque), EN→ES, EN→ZH
Lapata & Padó 2005	Semantic role labeling	EN→DE
Das & Petrov 2011	POS tagging	EN→DA, DE, EL, ES, IT, NL, PT, SV
McDonald et al. 2011	Dependency parsing	All combinations of EN, DA, DE, EL, ES, IT, NL, PT, SV
Duong et al. 2013	POS tagging	EN→DA, DE, EL, ES, IT, NL, PT, SV

Annotation projection



Vulić, Søgaard & Faruqui: Tutorial on cross-lingual word representations, EMNLP 2017.
http://people.ds.cam.ac.uk/iv250/tutorial/xlingrep-tutorial.pdf

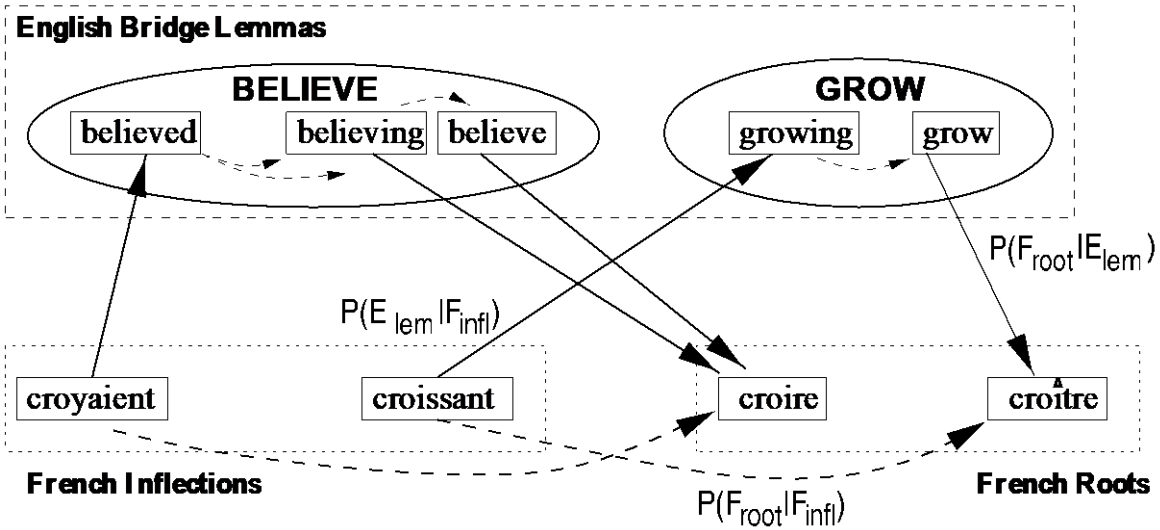
Morphology



- Arrows: alignment links
- Idea: *croyant* and *croire* belong to the same verb because they are linked by *believing*

Morphology

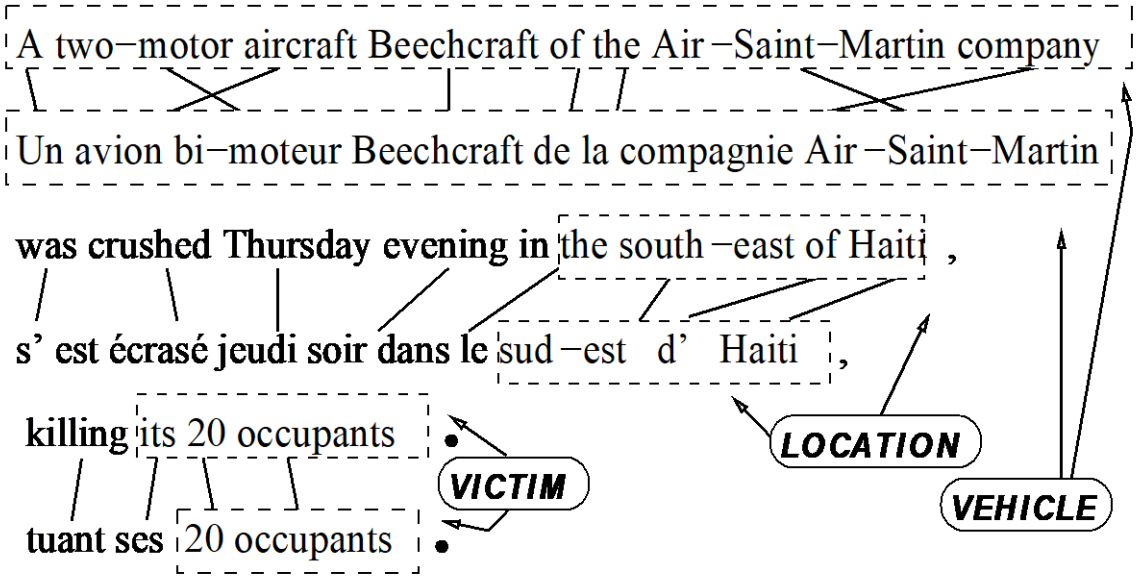
- The English side of the corpus can be annotated with lemmas:



  - *croyaient* is linked with *believed* (alignment)
  - *believed* is linked with *believe* (lemmatization)
  - *believe* is linked with *croire* (alignment)
  - *croyaient* is a morphological variant of *croire*

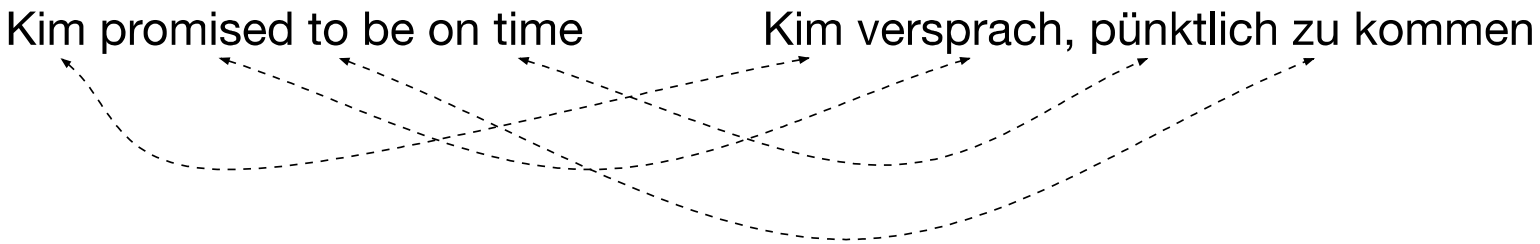
Information extraction

- Texts about plane crashes
- Ad-hoc annotation of the English data:
  - *location, vehicle, victim*
- Parallel data are created by MT system, thus noisy

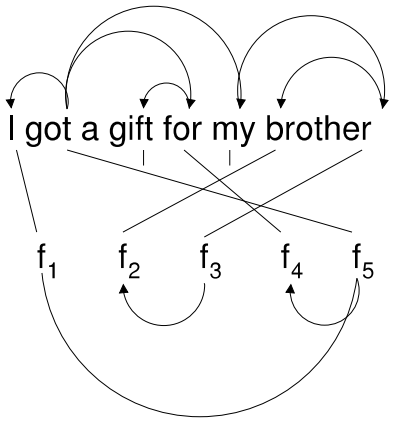


Semantic role labeling

- Both sides of the parallel corpus are parsed
- Role annotations are projected onto aligned constituents



Dependency parsing



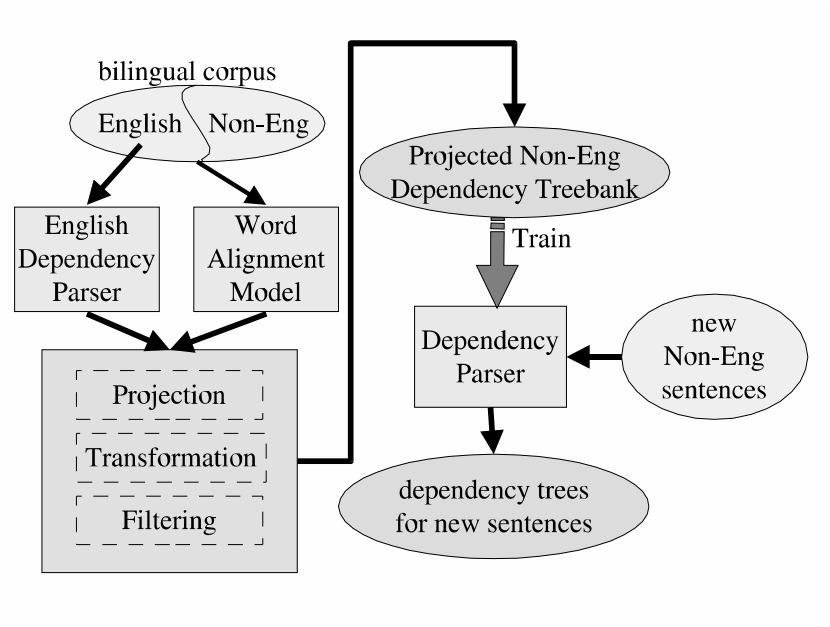
English dependencies

English sentence

Alignment

Foreign language sentence

Projected dependencies



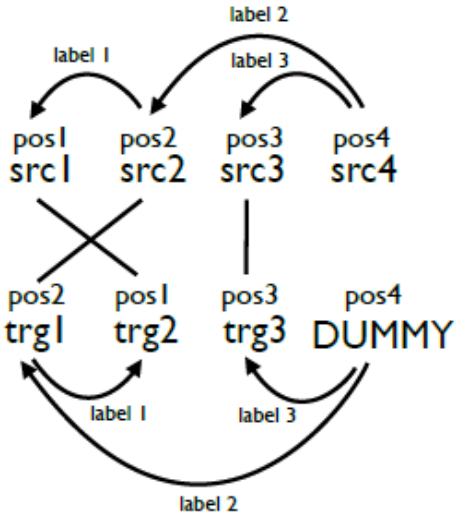
Direct Correspondence Assumption (DCA): Given a pair of sentences E and F that are (literal) translations of each other with syntactic structures $Tree_E$ and $Tree_F$, if nodes x_E and y_E of $Tree_E$ are aligned with nodes x_F and y_F of $Tree_F$, respectively, and if syntactic relationship $R(x_E, y_E)$ holds in $Tree_E$, then $R(x_F, y_F)$ holds in $Tree_F$.

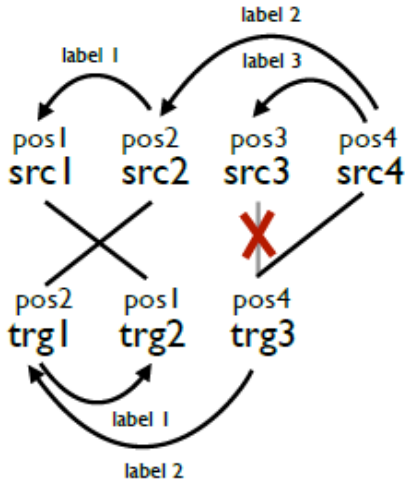
Dependency parsing

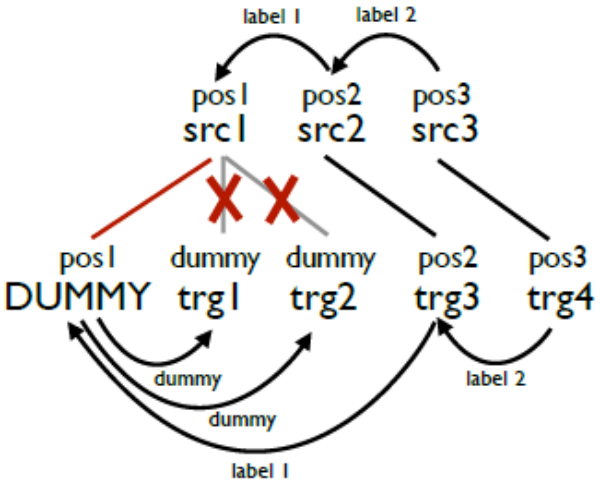
- DCA:
  - English-Spanish: 37% English-Chinese: 38%
- Main problem: Unaligned words
  - Aspect markers in Chinese
  - Reflexive pronouns in Spanish
- Solution: Post-projection transformation rules
  - Example: *An aspect marker modifies a verb to its left.*
- With transformation rules:
  - English-Spanish: 70% English-Chinese: 67%
- After training a target language parser on the projected data:
  - Spanish: 72% Chinese: 64%

Dependency parsing

- Tiedemann (2014): dummy nodes







Annotation projection

Authors	Task	Languages
Yarowsky & Ngai 2001	POS tagging NP chunking	EN→FR EN→FR, EN→ZH
Yarowsky et al. 2001	Morphology	EN→FR
Riloff et al. 2002	Information extraction	EN→FR
Diab & Resnik 2002	Word sense disambiguation	EN↔FR
Hwa et al. 2005	Dependency parsing	EN→EU (Basque), EN→ES, EN→ZH
Lapata & Padó 2005	Semantic role labeling	EN→DE
Das & Petrov 2011	POS tagging	EN→DA, DE, EL, ES, IT, NL, PT, SV
McDonald et al. 2011	Dependency parsing	All combinations of EN, DA, DE, EL, ES, IT, NL, PT, SV
Duong et al. 2013	POS tagging	EN→DA, DE, EL, ES, IT, NL, PT, SV

2011 – Was has changed?

- Parallel corpora are available for many language pairs
  - Europarl, OPUS, …
- POS-annotated texts and treebanks are available for many languages (with common label sets)
  - Required for evaluating the projected annotations (simulation of low-resource setting)
- More powerful machine learning techniques

- Google expands and is interested in language technology
  - Better search indexing for more languages
  - Parsing to improve machine translation and search
  - Parsing and coreference resolution for question answering systems

Annotation projection

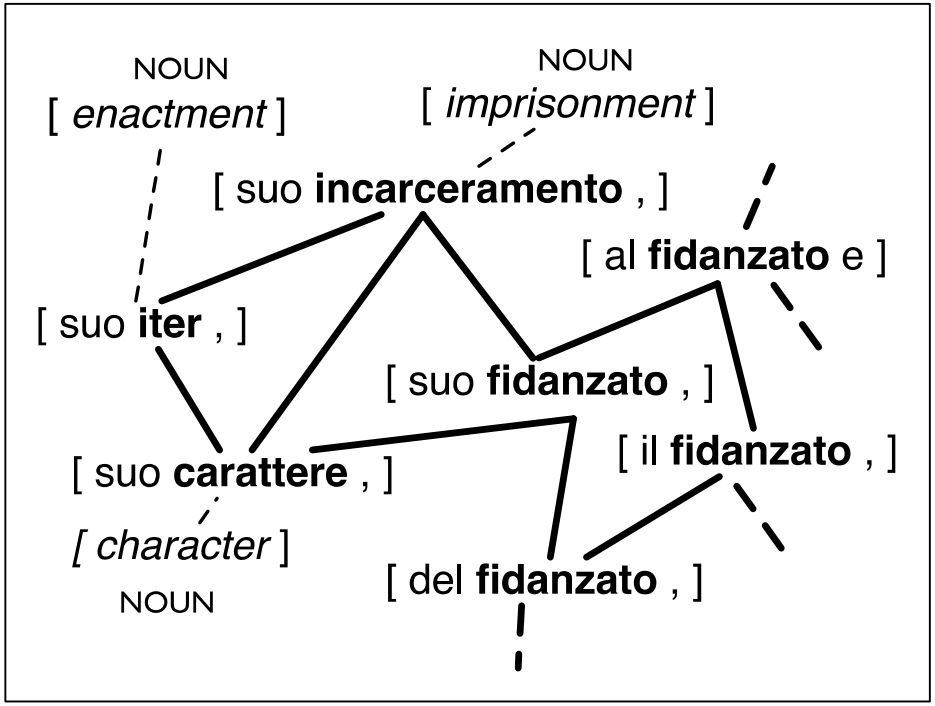
Authors	Task	Languages
Yarowsky & Ngai 2001	POS tagging NP chunking	EN→FR EN→FR, EN→ZH
Yarowsky et al. 2001	Morphology	EN→FR
Riloff et al. 2002	Information extraction	EN→FR
Diab & Resnik 2002	Word sense disambiguation	EN↔FR
Hwa et al. 2005	Dependency parsing	EN→EU (Basque), EN→ES, EN→ZH
Lapata & Padó 2005	Semantic role labeling	EN→DE
Das & Petrov 2011 ***Google***	POS tagging	EN→DA, DE, EL, ES, IT, NL, PT, SV
McDonald et al. 2011 ***Google***	Dependency parsing	All combinations of EN, DA, DE, EL, ES, IT, NL, PT, SV
Duong et al. 2013	POS tagging	EN→DA, DE, EL, ES, IT, NL, PT, SV

Das & Petrov 2011

- Improvements compared to Yarowsky & Ngai 2001:
  - No direct projection
    - Instead, create a graph and apply label propagation
  - Different machine learning technique for tagger training
    - Unsupervised instead of supervised learning
  - Universal POS tag set
    - 12 word categories aimed to be valid in all languages
    - Thus, easier comparison across languages

Das & Petrov 2011

- No direct projection, but graph creation and label propagation
  - Monolingual similarity
    - Links between words that occur in the same contexts
  - Bilingual similarity
    - Links between aligned words in the parallel corpus



Das & Petrov 2011

	Model	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Avg
baselines	EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
	Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
	Projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	78.8
our approach	No LP	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
	With LP	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	83.4
oracles	TB Dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*
	Supervised	*96.9*	*94.9*	*98.2*	*97.8*	*95.8*	*97.2*	*96.8*	*94.8*	*96.6*

- Projection: like Yarowsky & Ngai
- In all experiments, the source language is English

Das & Petrov 2011

	Model	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Avg
baselines	EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
	Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
	Projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	78.8
our approach	No LP	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
	With LP	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	83.4
oracles	TB Dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*
	Supervised	*96.9*	*94.9*	*98.2*	*97.8*	*95.8*	*97.2*	*96.8*	*94.8*	*96.6*

- No LP: tags are propagated only along one edge in the graph

Das & Petrov 2011

	Model	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	**Avg**
baselines	EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
	Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
	Projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	78.8
our approach	No LP	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
	With LP	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	83.4
oracles	TB Dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*
	Supervised	*96.9*	*94.9*	*98.2*	*97.8*	*95.8*	*97.2*	*96.8*	*94.8*	*96.6*

- With LP: with complete label propagation

Das & Petrov 2011

	Model	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Avg
baselines	EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
	Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
	Projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	78.8
our approach	No LP	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
	With LP	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	83.4
oracles	TB Dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*
	Supervised	*96.9*	*94.9*	*98.2*	*97.8*	*95.8*	*97.2*	*96.8*	*94.8*	*96.6*

- Supervised: high-resource scenario, for comparison purposes
  - Tagger trained directly with annotated data of target language

Duong et al. 2013

- The model of Das & Petrov is unnecessarily complicated
- Better results can be obtained with a simpler approach:
  1. Train a target language tagger using projected data (like Y&N)
  2. Improve the tagger by self-training and revision

- Projection – only with a small part of the parallel corpus:
  - Annotate source side
  - Align corpus
  - Project tags onto target side
    - Only use 1-1 alignment links
    - High precision, but many words remain untagged
  - Estimate emission and transition probabilities separately
    - Estimate transition probabilities only on the “best” sentences (< 10% untagged words, > 4 words long)

Duong et al. 2013

Algorithm 2 Self training and revision

1: Divide target language sentences into blocks of n sentences.
2: Tag the first block with the seed tagger.
3: Revise the tagged block.
4: Train a new tagger on the tagged block.
5: Add the previous tagger's lexicon to the new tagger.
6: Use the new tagger to tag the next block.
7: Goto 3 and repeat until all blocks are tagged.

Now, we use the rest of the parallel corpus...

- Step 3:
  - ws_i, ts_i: source language word and tag
  - wt_i, tt_i: target language source and tag
  - If ws_i is aligned with wt_j but $ts_i \neq tt_j$, then $tt_j \leftarrow ts_i$

Duong et al. 2013

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
Seed model	83.7	81.1	83.6	77.8	78.6	84.9	81.4	78.9	81.3
Self training + revision	**85.6**	**84.0**	**85.4**	80.4	81.4	86.3	83.3	**81.0**	83.4
Das and Petrov (2011)	83.2	79.5	82.8	**82.5**	**86.8**	**87.9**	**84.2**	80.5	**83.4**

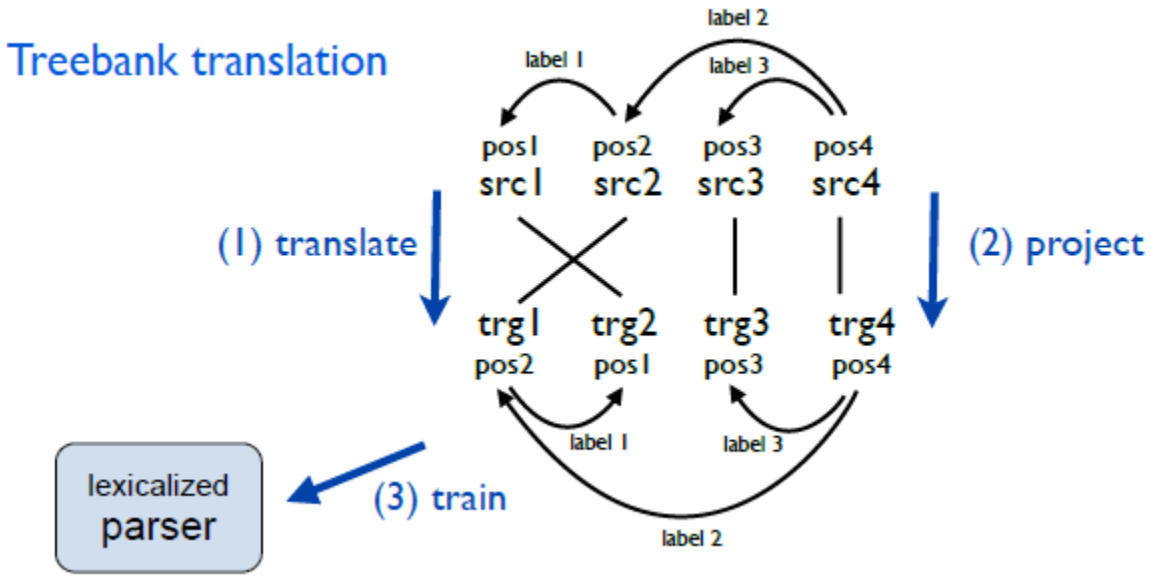
Training corpus translation

In case of parsing: treebank translation

Readings

- Jörg Tiedemann & Željko Agić (2016): *Synthetic treebanking for cross-lingual dependency parsing*. JAIR 55. (Sections 1 and 2)
  - What is the difference between *model transfer* and *data transfer*?
  - What is the difference between *annotation projection* and *treebank translation*? What are the advantages of the latter according to the authors?
  - What are the problems regarding cross-lingual parsing *evaluation*?

Treebank translation



Advantages

- Word alignment naturally comes out of the translation process and is not another source of noise.
- Work directly on the manually annotated data, not on automatically annotated data of language A, which may contain errors.
- (S)MT output tends to be more literal than manually translated sentences in a parallel corpus, so projection is more straigthtforward.

Drawbacks

- Depends on quality of MT system (and thus, on the available parallel corpora)
  - For closely related languages, rule-based (or otherwise bootstrapped) MT systems may be sufficient
- Depends on projection quality (but to the same extent as annotation projection)

Tomorrow...

Readings

- Ryan McDonald, Slav Petrov & Keith Hall (2011): *Multi-source transfer of delexicalized dependency parsers*. Proceedings of EMNLP. https://www.aclweb.org/anthology/D11-1006
- 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

Questions

- McDonald et al. 2011:
  - 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?

- Täckström et al. 2012:
  - Try to understand and explain Figure 1.