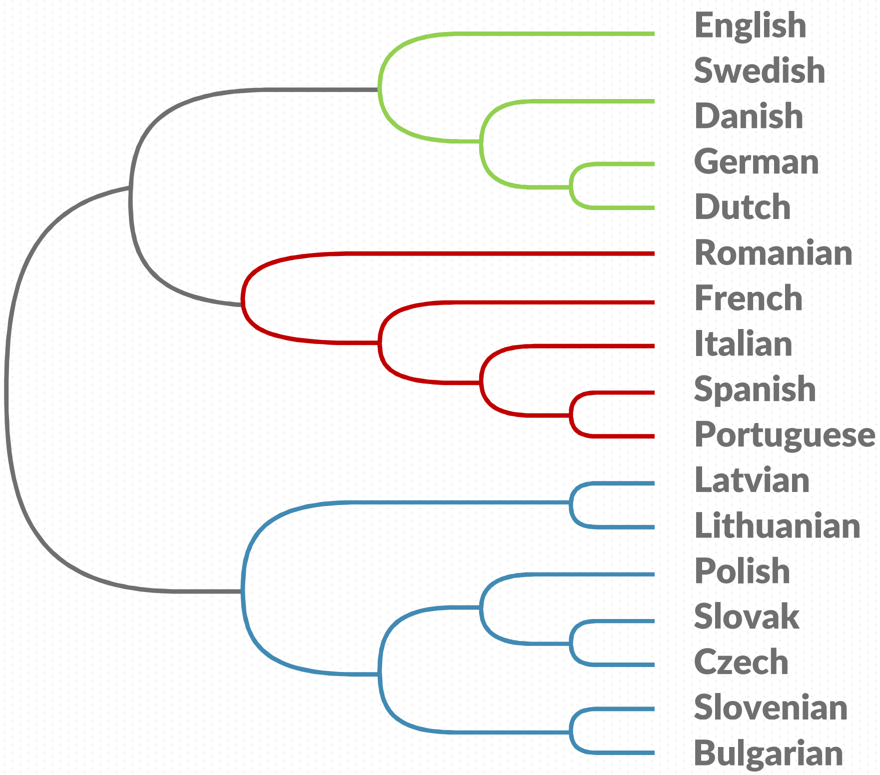
FOUND IN TRANSLATION:
Reconstructing Phylogenetic Language Trees from Translations

Ella Rabinovich[1,2], Noam Ordan[3], Shuly Wintner[2]
[1]IBM Research – Haifa, Israel
[2]Department of Computer Science, University of Haifa, Israel
[3]The Arab College for Education, Haifa, Israel

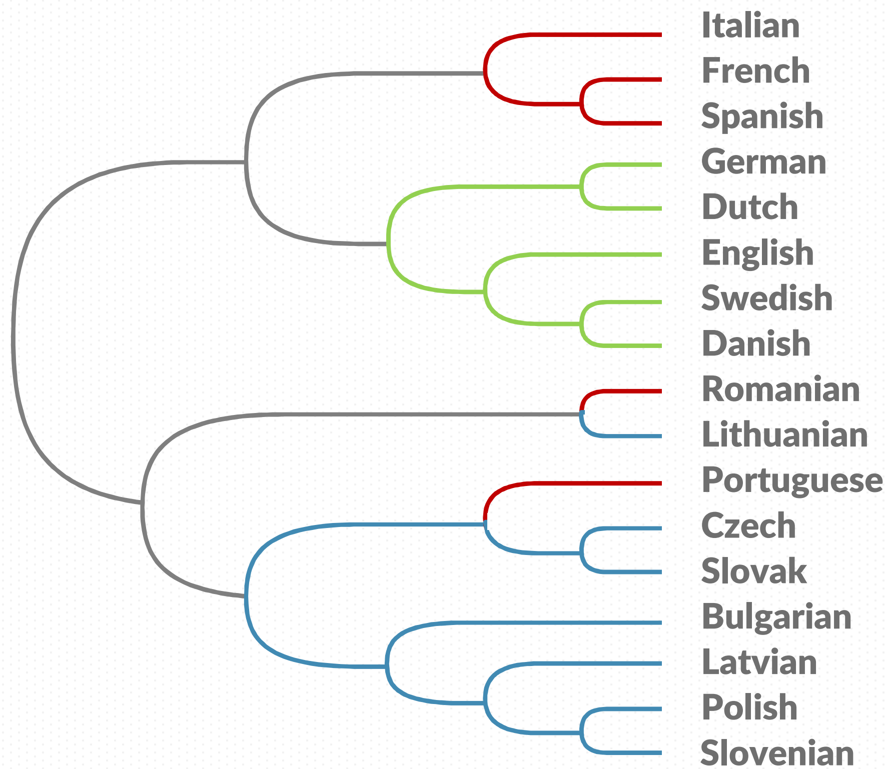
ACL 2017, Vancouver

STARTING FROM THE END (spoiler ☺)

the Indo-European phylogenetic tree
(the "ground truth")



phylogenetic tree reconstructed from
monolingual English texts translated from
17 IE languages



BACKGROUND – THE FEATURES OF TRANSLATIONESE

- Translators (almost) always tried to remain invisible
- Translations have unique characteristics that set them apart from originals
  - Universals (simplification, standardization, explicitation)
  - Interference (the "fingerprints" of a source language on the translation product)

HYPOTHESIS

Languages closer to each other **are likely to share more features** in the target language of translation

The distance between languages is retained and can be recovered when **assessed through these features in translated texts**

DATASET

- **Europarl (the proceedings of the European Parliament)**
  - Members are allowed to speak in any of the EU languages

- **All parliament speeches were translated from the original language into other EU languages using English as a pivot**
  - Direct translations into English, indirect translations into all other languages
  - We explore indirect translations into French in this work

---

- **We focus on 17 source languages, grouped into 3 language families**
  - Germanic, Romance, and Balto-Slavic

RECONSTRUCTION OF LANGUAGE TREES

FEATURES USED

- POS-trigrams, reflecting shallow syntactic structures (strongly associated with interference)
- Function words, reflecting grammar (associated with interference)
- Cohesive markers (associated with a translation universals)

AGGLOMERATIVE (HIERARCHICAL) CLUSTERING OF FEATURE VECTORS

- Using the variance minimization algorithm (Ward, 1963)

  → with Euclidean distance

IDENTIFICATION OF TRANSLATIONESE AND ITS SOURCE LANGUAGE

ORIGINAL VS. TRANSLATED

binary classification

Feature	English translations	French translations
POS-trigrams	97.60	98.40
function words	96.45	95.15
cohesive markers	86.50	85.25

CONFUSION MATRIX

source-language classification (POS-trigrams)

ENGLISH translations (76.5%)

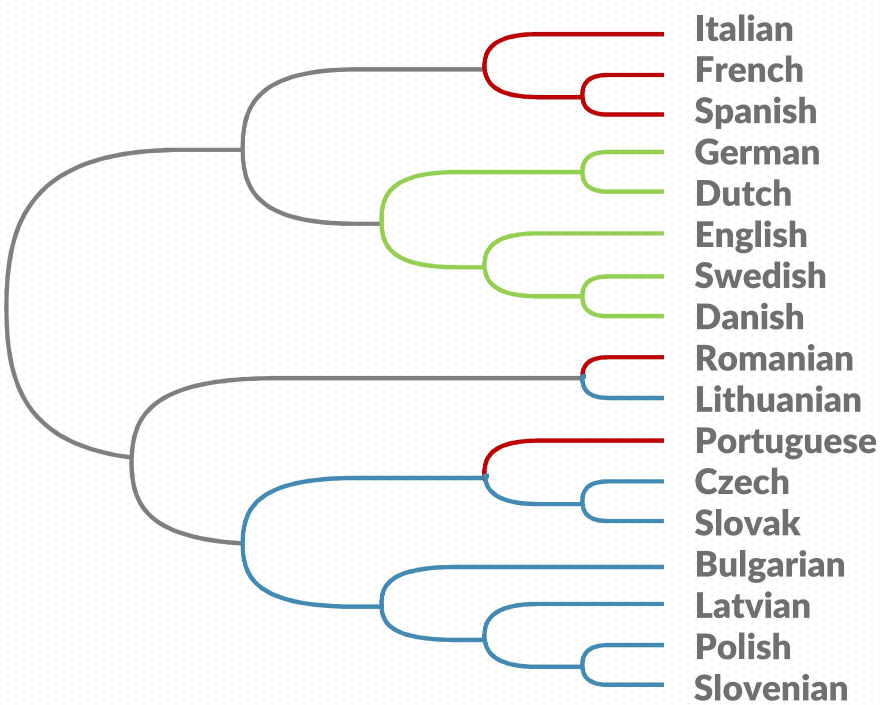
EN	NL	DE	DA	SV	PT	ES	FR	IT	RO	LT	PL	SK	CS	
84	4	2	2	4	0	0	1	1	0	0	0	1	1	EN
6	66	13	2	8	0	1	3	0	0	0	0	1	0	NL
2	16	71	2	2	0	3	4	0	0	0	0	0	0	DE
2	5	4	74	12	0	2	1	0	0	0	0	0	0	DA
4	4	1	13	73	0	0	4	1	0	0	0	0	0	SV
0	0	0	0	0	75	3	7	7	1	2	0	3	2	PT
1	0	2	2	1	3	74	11	5	0	0	0	0	1	ES
2	6	4	0	1	4	15	57	10	0	0	1	0	0	FR
3	0	4	0	0	13	4	12	63	0	0	0	0	1	IT
0	0	0	0	0	0	0	0	0	96	3	1	0	0	RO
0	0	0	0	0	1	0	0	0	2	93	0	3	1	LT
0	4	1	0	1	1	0	1	0	0	2	80	6	4	PL
2	0	0	1	0	1	0	0	0	0	1	5	78	10	SK
1	3	1	1	1	2	0	2	0	0	0	3	13	73	CS

FRENCH translations (48.9%)

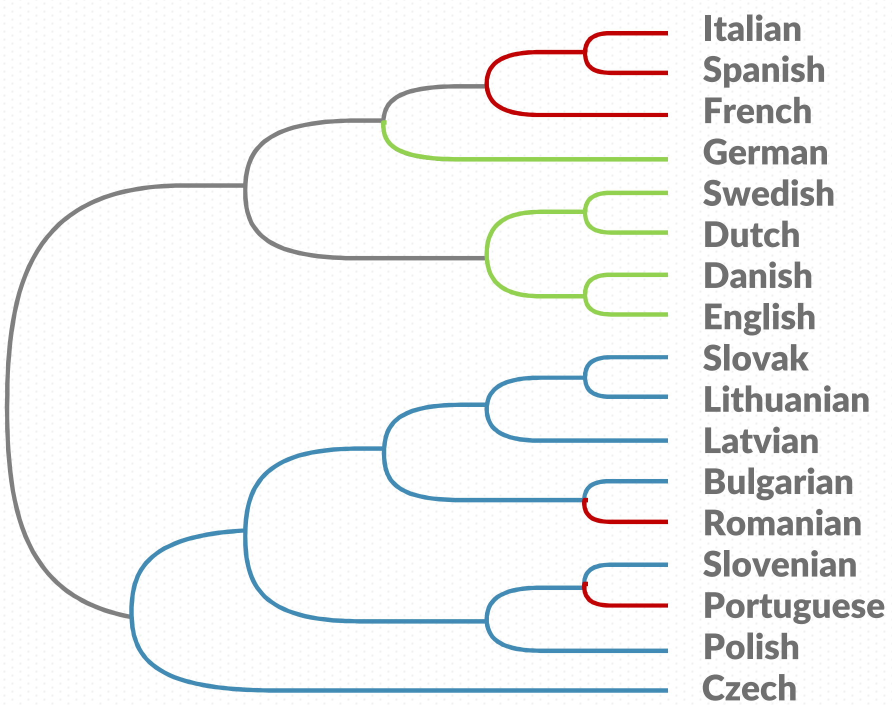
EN	NL	DE	DA	SV	PT	ES	FR	IT	RO	LT	PL	SK	CS	
40	11	3	7	7	1	6	2	2	0	0	9	6	6	EN
10	43	25	4	10	0	1	4	2	0	0	0	1	0	NL
8	32	41	1	5	0	6	3	1	0	0	1	2	0	DE
13	7	6	56	12	0	0	3	0	0	0	3	0	0	DA
7	13	9	5	46	2	3	1	2	0	0	6	2	4	SV
3	0	1	0	2	56	1	0	3	4	9	6	9	6	PT
4	4	4	0	3	3	54	7	15	0	0	2	3	1	ES
3	7	2	1	4	2	9	62	6	0	0	1	2	1	FR
4	0	4	0	8	9	18	11	41	0	0	3	1	1	IT
0	0	0	0	0	4	0	0	0	75	17	1	3	0	RO
1	0	0	0	0	12	0	0	0	22	54	2	4	5	LT
9	2	0	0	11	5	0	0	1	2	2	42	14	12	PL
4	1	0	1	1	10	0	1	3	2	8	14	38	17	SK
5	3	2	1	5	6	0	0	1	2	6	17	15	37	CS

RECONSTRUCTION OF LANGUAGE TREES

Phylogenetic language trees generated with translated text (POS-trigrams)



ENGLISH translations



FRENCH translations

EVALUATION METHODOLOGY

MEASURE SIMILARITY TO THE GOLD STANDARD

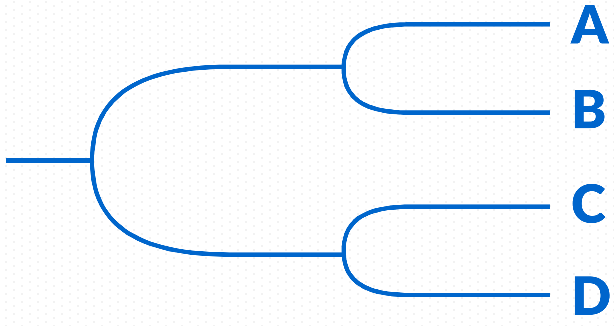
UNWEIGHTED EVALUATION
(CLADORGRAM)

assessing only structural (topological) similarity

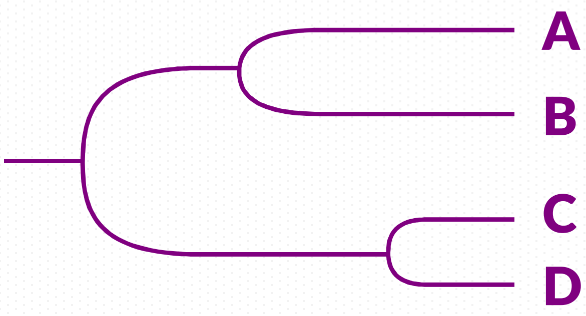
WEIGHTED EVALUATION
(PHYLOGRAM)

assessing similarity based on both structure and branching length

CLADOGRAM



PHYLOGRAM

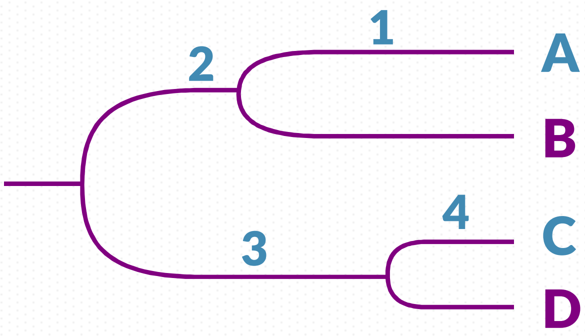


EVALUATION METHODOLOGY – CONT.

- **Adaptation of the L2-norm to leaf-pair distance**
- **Suitable for both weighted and unweighted evaluation**

$$Dist(t,g) = \sum_{i,j \in [1..N]; i \neq j} \left(D_t(l_i, lj) - D_g(l_i, lj) \right)^2$$

g – the gold tree
t – a tree subject to evaluation
$D_t(l_i, l_j)$ – distance between two leaves in a tree



EVALUATION RESULTS

DISTANCE OF A RECONSTRUCTED TREE FROM THE GOLD STANDARD
(using various feature sets)

UNWEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	.362	.07	**.367**	.06
POS-trigrams	**.353**	.06	.399	.08
Function words	.429	.07	.450	.08
Cohesive markers	.626	.16	.678	.14
Random tree	.724	.07	.724	.07

WEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	**.278**	.03	**.348**	.02
POS-trigrams	.301	.03	.351	.03
Function words	.304	.03	.376	.05
Cohesive markers	.598	.12	.636	.07
Random tree	.676	.10	.676	.10

trees built from **English** translations are systematically closer to the gold standard than trees built from translations into **French** (done via a third language)

the quality of trees increases for feature sets associated with **interference**

the worst tree is generated using cohesive markers

EVALUATION RESULTS

DISTANCE OF A RECONSTRUCTED TREE FROM THE GOLD STANDARD
(using various feature sets)

UNWEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	.362	.07	**.367**	.06
POS-trigrams	**.353**	.06	.399	.08
Function words	.429	.07	.450	.08
Cohesive markers	.626	.16	.678	.14
Random tree	.724	.07	.724	.07

WEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	**.278**	.03	**.348**	.02
POS-trigrams	.301	.03	.351	.03
Function words	.304	.03	.376	.05
Cohesive markers	.598	.12	.636	.07
Random tree	.676	.10	.676	.10

trees built from **English** translations are systematically closer to the gold standard than trees built from translations into **French** (done via a third language)

the quality of trees increases for feature sets associated with **interference**

the worst tree is generated using cohesive markers

EVALUATION RESULTS

DISTANCE OF A RECONSTRUCTED TREE FROM THE GOLD STANDARD
(using various feature sets)

UNWEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	.362	.07	**.367**	.06
POS-trigrams	**.353**	.06	.399	.08
Function words	.429	.07	.450	.08
Cohesive markers	.626	.16	.678	.14
Random tree	.724	.07	.724	.07

WEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	**.278**	.03	**.348**	.02
POS-trigrams	.301	.03	.351	.03
Function words	.304	.03	.376	.05
Cohesive markers	.598	.12	.636	.07
Random tree	.676	.10	.676	.10

trees built from **English** translations are systematically closer to the gold standard than trees built from translations into **French** (done via a third language)

the quality of trees increases for feature sets associated with **interference**

the worst tree is generated using cohesive markers

EVALUATION RESULTS

DISTANCE OF A RECONSTRUCTED TREE FROM THE GOLD STANDARD
(using various feature sets)

UNWEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	.362	.07	**.367**	.06
POS-trigrams	**.353**	.06	.399	.08
Function words	.429	.07	.450	.08
Cohesive markers	.626	.16	.678	.14
Random tree	.724	.07	.724	.07

WEIGHTED EVALUATION				
target language	English		French	
feature	AVG	STD	AVG	STD
POS-trigrams + FW	**.278**	.03	**.348**	.02
POS-trigrams	.301	.03	.351	.03
Function words	.304	.03	.376	.05
Cohesive markers	.598	.12	.636	.07
Random tree	.676	.10	.676	.10

trees built from **English** translations are systematically closer to the gold standard than trees built from translations into **French** (done via a third language)

the quality of trees increases for feature sets associated with **interference**

the worst tree is generated using cohesive markers

ANALYSIS

Articles

- Indefinite (“a”, “an”) and definite (“the”)

Possessive constructions

- With clitic ‘s (“the guest’s room”)
- With a prepositional phrase containing “of” (“the room of the guest”)
- With noun compounds (“guest room”)

Verb-particle constructions

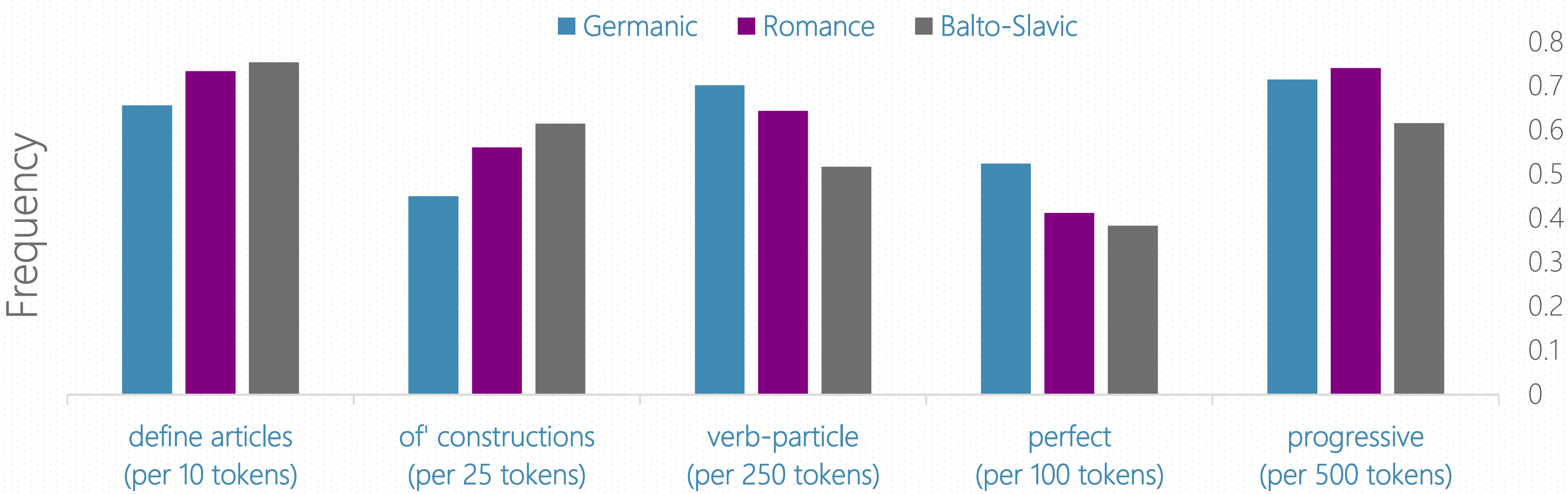
- Verbs that combine with a particle to create a new meaning (MWEs), e.g., “turn down”, “get over”

Tense and aspect

- With the auxiliary verbs “have” (present) or “be” (progressive), e.g., “have done”, “was going”

ANALYSIS – CONT.

FREQUENCIES reflecting various linguistic phenomena in English translations



SUMMARY

Translation does not distorts the original text randomly

A phylogenetic language tree can be reconstructed from monolingual texts translated from various languages

---

Features associated with interference (POS-ngrams, FWs) yield more accurate phylogenetic language trees

---

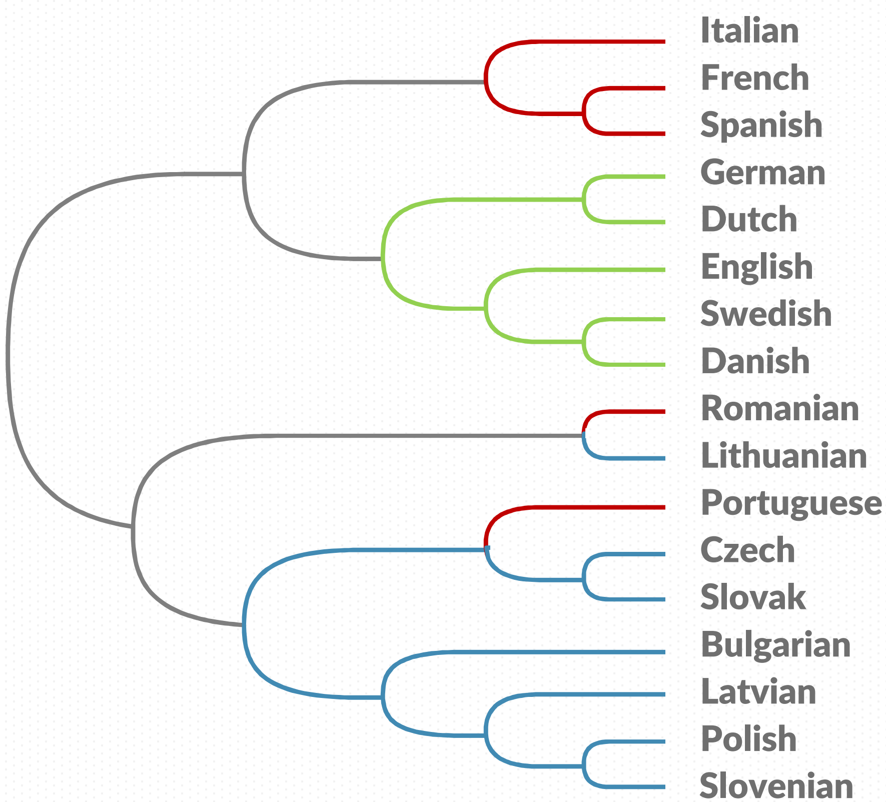
Translations impact the evolution of languages

- It is estimated that for certain languages up to 30% of published texts are mediated through translations (Pym and Chrupała, 2005)

Are translations likely to play a role in language change?

STARTING FROM THE END (spoiler ☺)

phylogenetic tree reconstructed from monolingual English texts translated from 17 IE languages



phylogenetic tree reconstructed from monolingual French texts translated indirectly from 17 IE languages via English pivot