On understanding and utilising the diversity of comparable corpora

Multilingual models

Serge Sharoff

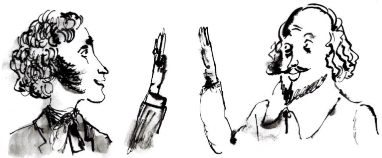
Centre for Translation Studies
University of Leeds

19 September 2022

Human needs → communication

The Merchant of Venice

If you prick us, do we not bleed?
If you tickle us, do we not laugh?
If you poison us, do we not die?
And if you wrong us, shall we not revenge?

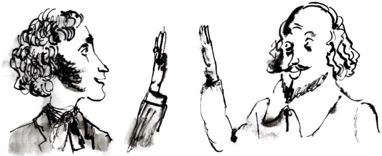


- Point of departure: we are all human...

Human needs → communication

The Merchant of Venice

If you prick us, do we not bleed?
If you tickle us, do we not laugh?
If you poison us, do we not die?
And if you wrong us, shall we not revenge?

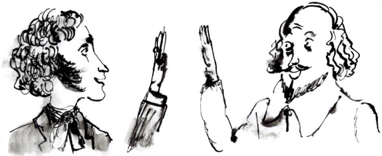


- Point of departure: we are all human...
- ...we share needs, desires, frustrations.

Human needs → communication

The Merchant of Venice

If you prick us, do we not bleed?
If you tickle us, do we not laugh?
If you poison us, do we not die?
And if you wrong us, shall we not revenge?



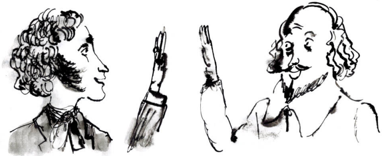
- Point of departure: we are all human. . .
- . . . we share needs, desires, frustrations.

→ We have developed many and varied means of expressing and negotiating these across different cultures, languages and kinds of language use

Human needs → communication

The Merchant of Venice

If you prick us, do we not bleed?
If you tickle us, do we not laugh?
If you poison us, do we not die?
And if you wrong us, shall we not revenge?



- Point of departure: we are all human. . .
- . . . we share needs, desires, frustrations.

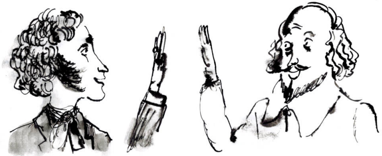
→ We have developed many and varied means of expressing and negotiating these across different cultures, languages and kinds of language use

- My needs concern better understanding of language

Human needs → communication

The Merchant of Venice

If you prick us, do we not bleed?
If you tickle us, do we not laugh?
If you poison us, do we not die?
And if you wrong us, shall we not revenge?



- Point of departure: we are all human...
- ...we share needs, desires, frustrations.

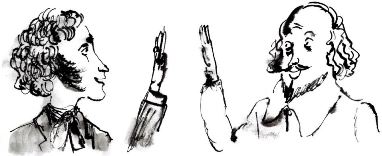
→ We have developed many and varied means of expressing and negotiating these across different cultures, languages and kinds of language use

- My needs concern better understanding of language
- My desires concern better tools to help language users

Human needs → communication

The Merchant of Venice

If you prick us, do we not bleed?
If you tickle us, do we not laugh?
If you poison us, do we not die?
And if you wrong us, shall we not revenge?



- Point of departure: we are all human. . .
- . . . we share needs, desires, frustrations.

→ We have developed many and varied means of expressing and negotiating these across different cultures, languages and kinds of language use

- My needs concern better understanding of language
- My desires concern better tools to help language users
- My frustrations concern under-resourced linguistic areas

Value of linguistic diversity

- In Ethnologue: 5,625 languages with $> 1,000$ speakers
  100 largest languages cover 85% world's population

Value of linguistic diversity

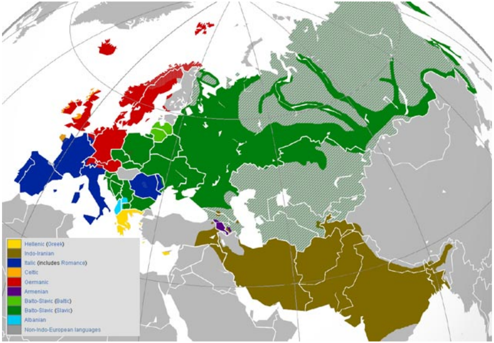
- In Ethnologue: 5,625 languages with $> 1,000$ speakers
  100 largest languages cover 85% world's population
- 98-100. Balochi, Belarusian and Konkani, $\approx$ 7M speakers

Value of linguistic diversity

- In Ethnologue: 5,625 languages with $> 1,000$ speakers
  100 largest languages cover 85% world's population
- 98-100. Balochi, Belarusian and Konkani, $\approx$ 7M speakers
- 40. Ukrainian, 30M native speakers (8. in Europe)

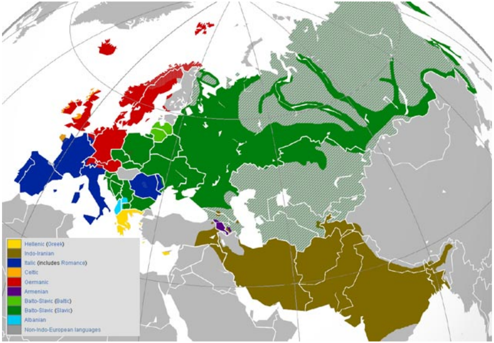
Value of linguistic diversity

- In Ethnologue: 5,625 languages with $> 1,000$ speakers
  100 largest languages cover 85% world's population
- 98-100. Balochi, Belarusian and Konkani, $\approx$ 7M speakers
- 40. Ukrainian, 30M native speakers (8. in Europe)



Value of linguistic diversity

- In Ethnologue: 5,625 languages with $> 1,000$ speakers
  100 largest languages cover 85% world's population
- 98-100. Balochi, Belarusian and Konkani, $\approx$ 7M speakers
- 40. Ukrainian, 30M native speakers (8. in Europe)



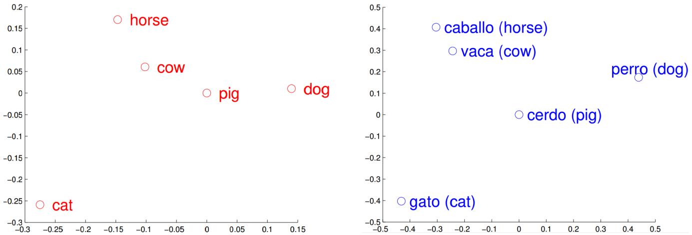
- {Czech, Russian} $\rightarrow$ {Belarusian, Ukrainian}
  {Hindi, Urdu} $\rightarrow$ {Konkani, Marathi}
  {Persian} $\rightarrow$ {Balochi, Kurdish}

Using similarities between languages

We need statistical data to predict parts of speech, gender, name categories, etc, for each language

English	Winston Churchill called the fall of Singapore the "worst disaster" in British history.
Czech	Winston Churchill označil pád Singapuru za „největší katastrofu“ v dějinách Británie.
Russian	Уинстон Черчилль назвал сдачу Сингапура «худшей катастрофой» в британской истории.
Belarusian	Уінстан Чэрчыль назваў падзенне Сінгапура "найгоршай катастрофай" у брытанскай гісторыі.
Polish	Winston Churchill nazwał sromotną klęskę Singapuru „najgorszą katastrofą” w historii Brytanii.
Ukrainian	Вінстон Черчилль назвав падіння Сінгапуру «найгіршою катастрофою» в британській історії.

Cross-lingual word embeddings (word2vec)



- Earlier vector models (Rapp, 1995)
- Predicting multi-word expressions (Sharoff et al., 2006)
- Linear transform or MLP for monolingual embeddings

$$\min_W \sum ||We_i - f_i||^2$$

- SGD (Mikolov et al., 2013), CCA (Faruqui and Dyer, 2014), multivariate regression (Dinu et al., 2014), regression with orthogonalisation constraints (Artetxe et al., 2016)

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al., 2013):

$$\min_{W} \sum ||We_i - f_i||^2$$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al., 2013):

$$\min_{W} \sum ||We_i - f_i||^2$$

- Orthogonality constraint (Artetxe et al., 2016):

$$W = V \times U^T$$

when V and U come from SVD factorisation of $F \times E^T$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al., 2013):

$$\min_{W} \sum ||We_i - f_i||^2$$

- Orthogonality constraint (Artetxe et al., 2016):

$$W = V \times U^T$$

  when V and U come from SVD factorisation of $F \times E^T$

- Adding Weighted Levenshtein Distance:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$$

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al., 2013):

$$\min_{W} \sum ||We_i - f_i||^2$$

- Orthogonality constraint (Artetxe et al., 2016):

$$W = V \times U^T$$

  when V and U come from SVD factorisation of $F \times E^T$

- Adding Weighted Levenshtein Distance:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1 - \alpha) WLD(s_e, s_f)$$

- Refinement for building cross-lingual spaces:

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al., 2013):

$$\min_{W} \sum ||We_i - f_i||^2$$

- Orthogonality constraint (Artetxe et al., 2016):

$$W = V \times U^T$$

  when V and U come from SVD factorisation of $F \times E^T$

- Adding Weighted Levenshtein Distance:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1-\alpha) WLD(s_e, s_f)$$

- Refinement for building cross-lingual spaces:
  1. Large dictionary of reliable cognates

Integration of WLD into embeddings

- Cross-lingual spaces (Mikolov et al., 2013):

$$\min_{W} \sum ||We_i - f_i||^2$$

- Orthogonality constraint (Artetxe et al., 2016):

$$W = V \times U^T$$

when V and U come from SVD factorisation of $F \times E^T$

- Adding Weighted Levenshtein Distance:

$$score(s_e, s_f) = \alpha \cos(v_e, v_f) + (1-\alpha) WLD(s_e, s_f)$$

- Refinement for building cross-lingual spaces:
  1. Large dictionary of reliable cognates
  2. Re-alignment of spaces using this dictionary

Dictionary induction results

State-of-the-art for en-it (Artetxe, et al 2016)	0.393
Weighted Levenshtein Distance	0.531

- When selecting cognates only (45%)
  This removes questionable translation equivalents:
  absolve / esimere or *abysmally / malo* ('bad(ly)')

State-of-the-art (Artetxe, et al 2016)	0.601
Weighted Levenshtein Distance	0.692

Dictionary induction results

State-of-the-art for en-it (Artetxe, et al 2016)	0.393
Weighted Levenshtein Distance	0.531

- When selecting cognates only (45%)
  This removes questionable translation equivalents:
  absolve / esimere or *abysmally / malo* ('bad(ly)')

State-of-the-art (Artetxe, et al 2016)	0.601
Weighted Levenshtein Distance	0.692

- Producing cross-lingual Panslavonic embeddings:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
SOTA:	0.429	0.611	0.584	0.566	0.929	0.814
With WLD:	0.840	0.763	0.751	0.662	0.945	0.910

Dictionary induction results

State-of-the-art for en-it (Artetxe, et al 2016)	0.393
Weighted Levenshtein Distance	0.531

- When selecting cognates only (45%)
  This removes questionable translation equivalents:
  absolve / *esimere* or *abysmally* / *malo* ('bad(ly)')

State-of-the-art (Artetxe, et al 2016)	0.601
Weighted Levenshtein Distance	0.692

- Producing cross-lingual Panslavonic embeddings:

	sl-hr	sl-cs	sl-pl	sl-ru	ru-uk	cs-sk
SOTA:	0.429	0.611	0.584	0.566	0.929	0.814
With WLD:	0.840	0.763	0.751	0.662	0.945	0.910

- In-family embedding spaces are better than multilingual ones:
  Success in NER Shared task at BSNLP'17 (Sharoff, 2020)

Multilingual contextual embeddings

Contextual embeddings: BERT-like models

*I put my glass on the kitchen **table**.* →
*J'ai posé mon verre sur la **table** de la cuisine.*
*The **table** lists all the products.* →
*Le **tableau** contient la liste de tous les produits.*

- Multilingual pre-trained embeddings align their representations: mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020)

Multilingual contextual embeddings

Contextual embeddings: BERT-like models

*I put my glass on the kitchen **table**.* $\rightarrow$
*J'ai posé mon verre sur la **table** de la cuisine.*
*The **table** lists all the products.* $\rightarrow$
*Le **tableau** contient la liste de tous les produits.*

- Multilingual pre-trained embeddings align their representations: mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020)
- Efficient zero-shot performance
  Training on English only $\rightarrow$ [Уинстон Черчилль]$_{\text{PER}}$

Multilingual contextual embeddings

Contextual embeddings: BERT-like models

*I put my glass on the kitchen **table**.* →
*J'ai posé mon verre sur la **table** de la cuisine.*
*The **table** lists all the products.* →
*Le **tableau** contient la liste de tous les produits.*

- Multilingual pre-trained embeddings align their representations: mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020)
- Efficient zero-shot performance
  Training on English only → [Уинстон Черчилль]$_{\mathrm{PER}}$

BUT we cannot use WLD to align *word* spaces, we need to finetune transformer parameters

Making synthetic corpora

- *Winston Churchill = Уінстан Чэрчыль = Вінстон Черчилль*
  Singapore = Сингапур = Сінгапур = Сінгапур

Making synthetic corpora

- *Winston Churchill = Уінстан Чэрчыль = Вінстон Черчилль*
  Singapore = Сингапур = Сінгапур = Сінгапур

→ Person: *Winston Churchill, Albert Einstein, Anastasia Romanova*
Location: *Singapore, Algeria, Anadyr, Brahmaputra, Ambracia, Zambia*

Making synthetic corpora

- *Winston Churchill = Уінстан Чэрчыль = Вінстон Черчилль*
  Singapore = Сингапур = Сінгапур = Сінгапур

→ Person: *Winston Churchill, Albert Einstein, Anastasia Romanova*
  Location: *Singapore, Algeria, Anadyr, Brahmaputra, Ambracia, Zambia*

- Case, Gender, Number as detected by UDPipe

Making synthetic corpora

- *Winston Churchill* = *Уінстан Чэрчыль* = *Вінстон Черчилль*
  Singapore = *Сингапур* = *Сінгапур* = *Сінгапур*

→ Person: *Winston Churchill, Albert Einstein, Anastasia Romanova*
  Location: *Singapore, Algeria, Anadyr, Brahmaputra, Ambracia, Zambia*

- Case, Gender, Number as detected by UDPipe
- Fine-tuning on a large annotated corpus of Russian, plus WikiMatrix templates (Schwenk et al., 2019) in under-resourced languages:
  Уінстан Чэрчыль назваў падзенне Сінгапура "найгоршай катастрофай" у брытанскай гісторыі.
  → Альберт Эйнштэйн назваў падзенне Амбракіі "найгоршай катастрофай" у брытанскай гісторыі.
  'Albert Einstein called the fall of Ambracia the "worst disaster" in British history.'

Predicting NER

Belarusian

	Zero-shot	MT	Synthetic
PER	0.88	0.89	0.92
LOC	0.64	0.67	0.76
ORG	0.54	0.56	0.61

Polish

	Zero-shot	MT	Synthetic
PER	0.87	0.89	0.92
LOC	0.68	0.70	0.80
ORG	0.43	0.48	0.66

ORG challenge: *Międzynarodowe Centrum Badań nad Ochroną i Konserwacją Dziedzictwa Kulturowego*
'International Centre for the Study of the Preservation and Restoration of Cultural Property'

Functional Text Dimensions (Sharoff, 2021)

news To what extent does the text provide an informative report of recent events? (Prototype: *newswires*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Functional Text Dimensions (Sharoff, 2021)

news To what extent does the text provide an informative report of recent events? (Prototype: *newswires*)

argum To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion columns*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Functional Text Dimensions (Sharoff, 2021)

news To what extent does the text provide an informative report of recent events? (Prototype: *newswires*)

argum To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion columns*)

review To what extent does the text evaluate a specific entity? (*reviews of products or locations*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Functional Text Dimensions (Sharoff, 2021)

news To what extent does the text provide an informative report of recent events? (Prototype: *newswires*)

argum To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion columns*)

review To what extent does the text evaluate a specific entity? (*reviews of products or locations*)

personal To what extent does the text report a personal story? (*diary-like blog entries*)

Rating for functions with respect to prototypes

0	none;	Ignore hesitations
0	slightly;	$\Uparrow$
.5	somewhat or partly;	$\Downarrow$
1	strongly.	Emphasise confident judgements

Principal Functional Text Dimensions

Code	Label	Prototypes	En	Ru
A1	argum	Argumentative blogs or opinion pieces	375	345
A8	news	Reporting newswire articles	207	538
A17	review	Reviews of products or experiences	102	257
A11	personal	Diary-like blog entries	161	284
A12	promotion	Adverts	350	331
A14	academic	Academic research papers	126	223
A16	info	Encyclopedic articles or textbooks	244	313
A7	instruct	Tutorials or FAQs	221	96
A9	legal	Laws, contracts, copyrights, T&Cs	95	105
A4	fiction	Fiction, myths, film plots	103	97
		Total training texts	1562	1930
A13	propaganda	Non-commercial promotion	73	62
A20	appell	Small ads, requests, CFPs	69	31

Prediction of text positions in this space and the nearest prototype.
Lots of methods → XLM-Roberta

Classification accuracy

FTD	F1		CI
Argument	0.729	±	0.021
News	0.944	±	0.011
Review	0.711	±	0.030
Personal	0.725	±	0.028
Promotion	0.937	±	0.012
Academic	0.883	±	0.023
Information	0.657	±	0.047
Instruction	0.760	±	0.104
Legal	0.757	±	0.039
Fiction	0.690	±	0.051

Confidence intervals from 10-fold cross-validation
Overall macro-F1 is 0.78± 0.037

Confusion matrix for classification

Predicted→ Reference↓	A1	A4	A7	A8	A9	A11	A12	A14	A16	A17
A1.argument	187	3	3	9	3	15	3	4	11	13
A4.fiction	6	50	0	0	0	12	0	0	2	6
A7.instruct	4	0	41	1	1	3	7	4	6	2
A8.news	13	0	1	446	2	3	2	1	8	4
A9.legal	4	0	0	0	59	2	1	2	5	0
A11.personal	13	3	0	2	0	134	3	0	4	18
A12.promotion	3	0	0	3	0	8	276	5	6	7
A14.academic	9	0	1	0	3	4	2	161	6	0
A16.info	19	0	0	11	4	10	1	10	90	6
A17.review	6	0	0	4	0	9	1	0	3	136

Composition of large Web corpora for pre-training

FTD	Wiki		OWT		CC-en		CC-ru	
Argument	0.88%	30720	57.11%	3635294	15.41%	2590419	8.85%	1945940
News	1.14%	39665	6.75%	429535	20.66%	3472767	13.12%	2884525
Review	4.68%	162511	3.74%	238156	6.03%	1012738	7.54%	1658324
Personal	0.03%	1168	20.27%	1290289	7.34%	1233399	11.34%	2492128
Promotion	0.07%	2390	2.67%	169988	15.07%	2533101	25.00%	5495815
Academic	0.82%	28558	2.51%	159921	3.41%	573081	3.11%	683447
Information	91.98%	3196502	1.18%	74886	15.34%	2577607	22.56%	4959489
Instruction	0.30%	10509	4.57%	290591	12.88%	2164862	2.24%	493385
Legal	0.04%	1340	1.16%	73620	2.14%	360195	0.41%	91124
Fiction	0.05%	1677	0.05%	3247	1.72%	289379	5.82%	1279432

Pre-trained transformer models (BERT, Roberta, GPT):

- BERT for English is trained by combining Wiki and fiction, mBERT is trained on Wikipedia for all languages
- OWT (used in GPT-2) comes from upvoted links on Reddit
- CC (used in XLM-Roberta) comes from Common Crawl

Explaining neural networks through linguistics

- Machine Learning predictions → reasons

Explaining neural networks through linguistics

- Machine Learning predictions $\rightarrow$ reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)

Explaining neural networks through linguistics

- Machine Learning predictions $\rightarrow$ reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :

Explaining neural networks through linguistics

- Machine Learning predictions $\rightarrow$ reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*

Explaining neural networks through linguistics

- Machine Learning predictions → reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*
  - public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny. . .*

Explaining neural networks through linguistics

- Machine Learning predictions → reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*
  - public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny. . .*
- Parts-of-speech: prepositions, past tense verbs, nominalisations (nouns ending in -*tion, ness, ment, ity*)

Explaining neural networks through linguistics

- Machine Learning predictions $\rightarrow$ reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*
  - public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny. . .*
- Parts-of-speech: prepositions, past tense verbs, nominalisations (nouns ending in *-tion, ness, ment, ity*)
- Syntactic: *be* as the main verb, by-passives, . . .

Explaining neural networks through linguistics

- Machine Learning predictions → reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*
  - public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny. . .*
- Parts-of-speech: prepositions, past tense verbs, nominalisations (nouns ending in -*tion, ness, ment, ity*)
- Syntactic: *be* as the main verb, by-passives, . . .
- Text-level: type-token ratio, word length, . . .

Explaining neural networks through linguistics

- Machine Learning predictions → reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*
  - public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny. . .*
- Parts-of-speech: prepositions, past tense verbs, nominalisations (nouns ending in *-tion, ness, ment, ity*)
- Syntactic: *be* as the main verb, by-passives, . . .
- Text-level: type-token ratio, word length, . . .
- Extraction for English, French, Russian and Spanish: passives with an agent, *do* as pro-verb

Explaining neural networks through linguistics

- Machine Learning predictions → reasons
- Lexicogrammatical features from (Biber, 1988; Sharoff, 2021)
- Lexical: publicVerbs, timeAdverbials, amplifiers, . . . :
  - amplifiers = *absolutely, altogether, completely, enormously. . .*
  - public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny. . .*
- Parts-of-speech: prepositions, past tense verbs, nominalisations (nouns ending in *-tion, ness, ment, ity*)
- Syntactic: *be* as the main verb, by-passives, . . .
- Text-level: type-token ratio, word length, . . .
- Extraction for English, French, Russian and Spanish: passives with an agent, *do* as pro-verb
- Training logistic regression on neural network predictions using interpretable linguistic features

Functions to features

Features	Promotion	Argum	News	Fiction	Personal	Academic
Type-Token Ratio	+ +		+ +	+ +	- - -	
Word length	+ +		+ + +	- -	-	
Adverbs	+ +		- -			- -
Conjunctions		+ +	+	-		
Discourse particles	- -			-	+ +	-
Nouns	+ +			+	- -	
Nominalisations		+ +		- -	- - -	+ +
Prepositions			+ +	-	-	-
Pronouns, 1p		- -	- - -	-	+ + +	- - -
Pronouns, 2p	+ +		- -	+		- - -
Pronouns, 3p	- -		- -	+ + +	- -	- -
Pronouns, WH-		+ +			-	
Verbs, past	- -		+ +	+ + +	+ +	- -
Verbs, present				+ +	+	+
Attributive adjectives	+ +	+ +	- -			
Negation	- -	+		+ +		- -
Subordinate clauses	+	+		-	-	

Features across functions in English

nominalisations (E14), nouns (E16), *by*-passives (F18), public verbs (K55) and clause negation (P67)

	E14		E16		F18		K55		P67	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Overall:	2.92%	2.46%	19.17%	18.96%	0.10%	0.00%	0.24%	0.00%	12.54%	7.30%
Arguing	3.29%	2.99%	17.90%	17.75%	0.10%	0.00%	0.36%	0.27%	17.58%	12.99%
Fiction	1.38%	1.19%	14.77%	14.57%	0.07%	0.00%	0.59%	0.46%	26.32%	17.10%
Instruct	2.73%	2.32%	19.48%	19.36%	0.08%	0.00%	0.21%	0.00%	16.04%	10.69%
News	3.20%	2.97%	18.39%	18.18%	0.13%	0.00%	0.55%	0.43%	9.11%	4.16%
Legal	5.36%	5.16%	19.65%	19.56%	0.18%	0.12%	0.29%	0.19%	21.82%	15.75%
Personal	1.66%	1.39%	16.73%	16.56%	0.06%	0.00%	0.33%	0.23%	13.99%	9.92%
Promoting	3.42%	3.03%	21.03%	20.95%	0.08%	0.00%	0.14%	0.00%	8.60%	0.00%
Academic	4.28%	3.99%	20.39%	20.35%	0.13%	0.00%	0.17%	0.03%	8.90%	0.00%
Inform	2.50%	2.07%	17.87%	17.66%	0.15%	0.00%	0.15%	0.00%	8.11%	0.00%
Review	1.76%	1.59%	17.56%	17.50%	0.07%	0.00%	0.26%	0.14%	13.15%	9.38%

Features across functions in Russian

nominalisations (E14), nouns (E16), *by*-passives (F18), public verbs (K55) and clause negation (P67)

	E14 English	E14 Russian	E16 English	E16 Russian	F18 English	F18 Russian	K55 English	K55 Russian	P67 English	P67 Russian
Overall:	2.46%	5.46%	18.96%	21.42%	0.00%	0.15%	0.00%	0.00%	7.30%	7.68%
Arguing	2.99%	5.47%	17.75%	19.41%	0.00%	0.15%	0.27%	0.10%	12.99%	11.73%
Fiction	1.19%	2.18%	14.57%	18.21%	0.00%	0.06%	0.46%	0.16%	17.10%	14.78%
Instruct	2.32%	4.48%	19.36%	21.11%	0.00%	0.11%	0.00%	0.00%	10.69%	11.60%
News	2.97%	6.16%	18.18%	22.22%	0.00%	0.00%	0.43%	0.00%	4.16%	3.38%
Legal	5.16%	11.07%	19.56%	22.41%	0.12%	0.59%	0.19%	0.00%	15.75%	5.04%
Personal	1.39%	2.80%	16.56%	18.19%	0.00%	0.00%	0.23%	0.07%	9.92%	13.33%
Promoting	3.03%	6.46%	20.95%	22.64%	0.00%	0.17%	0.00%	0.00%	0.00%	5.13%
Academic	3.99%	9.80%	20.35%	22.93%	0.00%	0.31%	0.03%	0.00%	0.00%	4.29%
Inform	2.07%	5.78%	17.66%	22.27%	0.00%	0.27%	0.00%	0.00%	0.00%	5.96%
Review	1.59%	3.59%	17.50%	19.39%	0.00%	0.00%	0.14%	0.00%	9.38%	10.81%

Take-home message

- Communication is reasonably universal across languages

→ we can create multilingual models

- Weighted Levenshtein Distance is efficient for better alignment of word embedding spaces across related languages
- We can build better models for under-resourced languages from natural annotation, i.e. existing annotations created for a different "natural" purpose
- Efficient development of models for Slavic NER from templates
- Corpora (even when collected using comparable methods) vary with respect to their functions

→ This impacts the frequencies of various features

References I

Artetxe, M., Labaka, G., and Agirre, E. (2016).
Learning principled bilingual mappings of word embeddings while preserving monolingual invariance.
In *Proc EMNLP*, Austin, Texas.

Biber, D. (1988).
Variation Across Speech and Writing.
Cambridge University Press.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020).
Unsupervised cross-lingual representation learning at scale.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

References II

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018).
BERT: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805.

Dinu, G., Lazaridou, A., and Baroni, M. (2014).
Improving zero-shot learning by mitigating the hubness problem.
arXiv preprint arXiv:1412.6568.

Faruqui, M. and Dyer, C. (2014).
Improving vector space word representations using multilingual correlation.
In *Proc EACL*, pages 462–471, Gothenburg, Sweden.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013).
Exploiting similarities among languages for machine translation.
arXiv preprint arXiv:1309.4168.

References III

Rapp, R. (1995).
Identifying word translations in non-parallel texts.
In *Proc. of the 33rd ACL*, pages 320–322, Cambridge, MA.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019).
WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia.
arXiv preprint arXiv:1907.05791.

Sharoff, S. (2020).
Finding next of kin: Cross-lingual embedding spaces for related languages.
Journal of Natural Language Engineering, 26:163–182.

Sharoff, S. (2021).
Genre annotation for the web: text-external and text-internal perspectives.
Register studies, 3:1–32.

References IV

Sharoff, S., Babych, B., and Hartley, A. (2006).
Using comparable corpora to solve problems difficult for human translators.
In *Proc International Confenrence on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006*, pages 739–746, Sydney.