# Word-order typology in Multilingual BERT: A case study in subordinate-clause detection

Dmitry Nikolaev, Sebastian Padó

SIGTYP 2022





#### Why subordinate-clause detection

The distinction between main and subordinate clauses is arguably universal *conceptually*: subordinate clauses (prototypically) encode *non-asserted information*.<sup>†</sup>

Crucial for claim, stance, and factivity detection.

#### Why subordinate-clause detection

However, it is highly variable on the level of superficial syntax. Therefore

- Different strategies needed for different languages
- Especially difficult in the zero-shot setting
- $\rightarrow$  A useful window into the capabilities of cross-lingual models.

# The specific research question

How well can Transformer-based (more precisely, BERT-based) models solve this task and what determines their performance?

## Two settings

**Single-language**: fine-tune and test a pretrained single-language BERT model on 20 typologically diverse languages.

- Check the overall performance level and the influence of the model and training-set size
- Perform qualitative evaluation

# Two settings

**Single-language**: fine-tune and test a pretrained single-language BERT model on 20 typologically diverse languages.

- Check the overall performance level and the influence of the model and training-set size
- Perform qualitative evaluation

**Cross-lingual**: fine-tune mBERT on data from 10 languages, test in a many-to-many way on 27 languages.

- Look for patterns
- Perform qualitative evaluation

#### Data

Universal Dependencies corpora in different languages.

The Parallel Universal Dependencies collection was used as much as possible as test sets for comparability.

Main clauses: elements with the root tag.

Subordinate clauses: elements marked as acl, ccomp, advcl, csubj, and xcomp.

#### Single-language results (ordered by accuracy)

Language Mandarin Accuracy 88.7	Vietnamese 90	Korean 90.4		Hindi 93.6	German 94.1	Armenian 94.3	Turkish 95.1	Welsh 95.6	
Language Basque Accuracy 96.9	Spanish 97.1	Irish 97.4	U	Hebrew 98.2	Afrikaans 98.8	French 99	Japanese 99.1	Czech 99.6	

## Single-language results

Language Mandarin	Nietnamese	Korean	Arabic	Hindi	German	Armenian	Turkish	Welsh	Indonesian
Accuracy 88.7	90	90.4	91.2	93.6	94.1	94.3	95.1	95.6	96
Language Basque Accuracy 96.9	Spanish 97.1	Irish 97.4	English 97.9	Hebrew 98.2	Afrikaans 98.8	French 99	Japanese 99.1	Czech 99.6	

A lot of POS lability: the same words can act as main-clause predicates and subordinate-clause markers. Construction-heavy grammar.

## Single-language results

Language Mandarin Accuracy 88.7	Vietnamese 90	sector and the sector of the s		Hindi 93.6	German 94.1	Armenian 94.3	Turkish 95.1	Welsh 95.6	
Language Basque	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	
Accuracy 96.9	97.1	97.4	97.9	98.2	98.8	99	99.1	99.6	

About sixty clause connectors; many allow for both coordinative and subordinative readings, which changes the status of the first clause.

Single	-langi	uage re	esult			annotatio	ostly due onal discre the train a	epancie		
Language M		Vietnamese	Korean	Arabic	Hindi	German	Armenian	Turkish	Welsh	Indonesian
Accuracy 8		90	90.4	91.2	93.6	94.1	94.3	95.1	95.6	96
Language I	Basque	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	Russian
Accuracy 9		97.1	97.4	97.9	98.2	98.8	99	99.1	99.6	99.7

For languages with stable complementisers, the task is solved even with small models and very small training sets.

Target

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud
English	96	95	93	94	94	86	98	98	96	96	96	93	93	95	96	98
Russian	95	94	86	93	95	90	94	96	95	<b>97</b>	99	94	94	93	96	96
Czech	94	94	83	95	100	92	92	93	94	94	98	95	88	89	92	94
French	94	91	84	92	95	90	92	97	95	95	99	94	87	93	96	96
German	94	87	95	94	88	82	90	95	94	92	93	91	90	89	95	96
Arabic	90	96	76	85	84	90	85	87	86	85	84	87	84	85	89	85
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87
Turkish	67	61	61	65	71	<u>62</u>	64	69	75	77	71	73	73	68	68	74
Korean	51	53	63	53	61	52	51	54	59	65	59	59	57	55	54	61
Japanese	55	56	<u>41</u>	<u>39</u>	<u>52</u>	63	51	<u>52</u>	52	54	55	54	<u>58</u>	53	<u>51</u>	54
	pl	pt	ru	ru	SV	eu	hi	tr	ja	ja	ko	vi	th	zh	zh	
	pud	pud	pud	syntag	pud	bdt	pud	pud	gsd	pud	pud	vtb	pud	gsd	pud	mean
English	95	97	93	93	96	88	87	83	66	70	67	82	84	67	71	88.9
Russian	98	95	100	99	95	88	90	90	68	72	70	79	80	64	69	89.2
Czech	97	93	94	96	92	87	88	88	64	66	68	78	79	65	71	87.5
-							00	00	<u>v</u> .							
French	98	96	96	97	94	85	89	86	54	61	66	77	76	63	69	87.0
German	<b>98</b> 89	96 94									66 57	77 78	76 78		69 68	87.0 85.2
			96	97	94	85	89	86	54	61	<u>57</u>			<u>63</u>		
German	89	94	96 93	97 88	94 <b>97</b>	85 81	<b>89</b> 86	86 78	<u>54</u> 59	<u>61</u> <u>62</u>		78	78	<u>63</u> <u>67</u>	68	85.2
German Arabic	89 85	94 86	96 93 88	97 88 85	94 <b>97</b> 89	85 81 71	<b>89</b> 86 70	86 78 65	<u>54</u> <u>59</u> <u>63</u>	<u>61</u> <u>62</u> <u>66</u>	<u>57</u> <u>59</u>	78 74	78 79	<u>63</u> <u>67</u> <u>66</u>	68 <u>65</u>	85.2 80.3
German Arabic Mandarin	89 85 85	94 86 85	96 93 88 86	97 88 85 85	94 <b>97</b> 89 86	85 81 71 82	<b>89</b> 86 70 87	86 78 65 89	<u>54</u> <u>59</u> <u>63</u> 80	<u>61</u> <u>62</u> <u>66</u> 78	<u>57</u> <u>59</u> 77	78 74 74	78 79 80	<u>63</u> <u>67</u> <u>66</u> <b>91</b>	68 <u>65</u> <b>86</b>	85.2 80.3 84.6

12

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud
English	96	95	93	94	94	86	98	98	96	96	96	93	93	95	96	98
Russian	95	94	86	93	95	90	94	96	95	97	99	94	94	93	96	96
Czech	94	94	83	95	100	92	92	93	94	94	98	95	88	89	92	94
French	94	91	84	92	95	90	92	97	95	95	99	94	87	93	96	96
German	94	87	95	94	88	82	90	95	94	92	93	91	90	89	95	96
Arabic	90	96	76	85	84	90	85	87	86	85	84	87	84	85	89	85
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87
Turkish	67	61	61	65	71	<u>62</u>	64	69	75	77	71	73	73	68	68	74
Korean	51	<u>53</u>	63	53	61	<u>52</u>	51	<u>54</u>	<u>59</u>	65	<u>59</u>	59	57	55	<u>54</u>	61
Japanese	55	56	<u>41</u>	<u>39</u>	<u>52</u>	63	<u>51</u>	<u>52</u>	<u>52</u>	54	55	<u>54</u>	<u>58</u>	53	<u>51</u>	<u>54</u>
	pl	pt	ru	ru	SV	eu	hi	tr	ja	ja	ko	vi	th	zh	zh	maan
	pud	pud	pud	syntag	pud	bdt	pud	pud	and	pud	pud	vtb	pud	acd	pud	mean
English	ID PROPERTY.		-		-		puu	puu	gsd	puu	puu	100	puu	gsd	puu	
	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	95 <b>98</b>	<b>97</b> 95			-		-			L	-					88.9 <b>89.2</b>
			93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	
Russian	98	95	93 100	93 <b>99</b>	96 95	88 88	87 90	83 90	<u>66</u> <u>68</u>	70 72	67 70	<b>82</b> 79	<b>84</b> 80	<u>67</u> <u>64</u>	71 69	89.2
Russian Czech	<b>98</b> 97	95 93	93 <b>100</b> 94	93 <b>99</b> 96	96 95 92	<b>88</b> <b>88</b> 87	87 90 88	83 90 88	<u>66</u> <u>68</u> <u>64</u>	$     \frac{70}{72}     \underline{66} $	67 70 68	<b>82</b> 79 78	<b>84</b> 80 79	$\frac{\underline{67}}{\underline{64}}$	71 69 71	<b>89.2</b> 87.5
Russian Czech French	<b>98</b> 97 <b>98</b>	95 93 96	93 <b>100</b> 94 96	93 <b>99</b> 96 97	96 95 92 94	<b>88</b> <b>88</b> 87 85	87 90 88 <b>89</b>	83 90 88 86	<u>66</u> <u>68</u> <u>64</u> <u>54</u>	$     \frac{70}{72}     \underline{66}     \underline{61}   $	67 70 68 66	<b>82</b> 79 78 77	<b>84</b> 80 79 76	$     \begin{array}{r} \underline{67} \\ \underline{64} \\ \underline{65} \\ \underline{63} \end{array} $	71 69 71 69	<b>89.2</b> 87.5 87.0
Russian Czech French German	<b>98</b> 97 <b>98</b> 89	95 93 96 94	93 <b>100</b> 94 96 93	93 <b>99</b> 96 97 88	96 95 92 94 <b>97</b>	<b>88</b> <b>88</b> 87 85 81	87 90 88 <b>89</b> 86	83 90 88 86 78	<u>66</u> <u>68</u> <u>64</u> <u>54</u> <u>59</u>	$     \frac{70}{72} \\     \frac{66}{61} \\     \underline{62}   $	67 70 68 66 <u>57</u>	<b>82</b> 79 78 77 78	<b>84</b> 80 79 76 78	$     \begin{array}{r} \underline{67} \\ \underline{64} \\ \underline{65} \\ \underline{63} \\ \underline{67} \end{array} $	71 69 71 69 68	<b>89.2</b> 87.5 87.0 85.2
Russian Czech French German Arabic	<b>98</b> 97 <b>98</b> 89 85	95 93 96 94 86	93 <b>100</b> 94 96 93 88	93 <b>99</b> 96 97 88 85	96 95 92 94 <b>97</b> 89	<b>88</b> <b>88</b> 87 85 81 71	87 90 88 <b>89</b> 86 70	83 90 88 86 78 65		$     \frac{70}{72}     \frac{66}{61}     \frac{61}{62}     \frac{66}{66} $	$ \begin{array}{r} 67\\ 70\\ 68\\ 66\\ \underline{57}\\ \underline{59}\\ \end{array} $	<b>82</b> 79 78 77 78 78 74	<b>84</b> 80 79 76 78 79		71 69 71 69 68 <u>65</u>	<b>89.2</b> 87.5 87.0 85.2 80.3
Russian Czech French German Arabic Mandarin	<b>98</b> 97 <b>98</b> 89 85 85	95 93 96 94 86 85	93 <b>100</b> 94 96 93 88 88	93 <b>99</b> 96 97 88 85 85	96 95 92 94 <b>97</b> 89 86	<b>88</b> <b>88</b> 87 85 81 71 82	87 90 88 <b>89</b> 86 70 87	83 90 88 86 78 65 89	$     \begin{array}{r}             \underline{66} \\             \underline{68} \\             \underline{64} \\             \underline{54} \\             \underline{59} \\             \underline{63} \\             80         \end{array}     $	$     \frac{70}{72} \\     \frac{66}{61} \\     \frac{62}{66} \\     \frac{66}{78} $	$ \begin{array}{r} 67\\ 70\\ 68\\ 66\\ \underline{57}\\ \underline{59}\\ 77\\ \end{array} $	<b>82</b> 79 78 77 78 74 74	<b>84</b> 80 79 76 78 79 80	67 64 65 63 67 66 <b>91</b>	71 69 71 69 68 <u>65</u> <b>86</b>	<b>89.2</b> 87.5 87.0 85.2 80.3 84.6

Russian as the source language has the best overall perf.

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud
English	96	95	93	94	94	86	98	98	96	96	96	93	93	95	96	98
Russian	95	94	86	93	95	90	94	96	95	97	99	94	94	93	96	96
Czech	94	94	83	95	100	92	92	93	94	94	98	95	88	89	92	94
French	94	91	84	92	95	90	92	97	95	95	99	94	87	93	96	96
German	94	87	95	94	88	82	90	95	94	92	93	91	90	89	95	96
Arabic	90	96	76	85	84	90	85	87	86	85	84	87	84	85	89	85
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87
Turkish	67	61	61	65	71	<u>62</u>	64	69	75	77	71	73	73	68	68	74
Korean	51	53	63	<u>53</u>	61	<u>52</u>	51	<u>54</u>	<u>59</u>	65	<u>59</u>	59	57	55	<u>54</u>	61
Japanese	55	56	<u>41</u>	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	55	<u>54</u>	<u>58</u>	53	<u>51</u>	<u>54</u>
	pl	pt	ru	ru	SV	eu	hi	tr	ja	ja	ko	vi	th	zh	zh	moon
	pud	pud	pud	syntag	pud	bdt	pud	pud	and a	nud	pud	vtb	pud	gsd	pud	mean
		puu	puu	Syntag	puu	Dut	puu	puu	gsd	pud	puu	100	pau	gou	puu	
English	95	97	93	93	96	88	87	83	<b>gsa</b>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
English Russian	95 <b>98</b>	1	-				-					_				88.9 <b>89.2</b>
-		97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	
Russian	98	<b>97</b> 95	93 100	93 <b>99</b>	96 95	88 88	87 90	83 90	<u>66</u> <u>68</u>	<u>70</u> 72	67 70	<b>82</b> 79	<b>84</b> 80	<u>67</u> <u>64</u>	71 69	89.2
Czech	<b>98</b> 97	<b>97</b> 95 93	93 <b>100</b> 94	93 <b>99</b> 96	96 95 92	<b>88</b> <b>88</b> 87	87 90 88	83 90 88	<u>66</u> <u>68</u> <u>64</u>	<u>70</u> 72 <u>66</u>	67 70 68	<b>82</b> 79 78	<b>84</b> 80 79	<u>67</u> <u>64</u> <u>65</u>	71 69 71	<b>89.2</b> 87.5
Russian Czech French	<b>98</b> 97 <b>98</b>	<b>97</b> 95 93 96	93 <b>100</b> 94 96	93 <b>99</b> 96 97	96 95 92 94	<b>88</b> <b>88</b> 87 85	87 90 88 <b>89</b>	83 90 88 86	<u>66</u> <u>68</u> <u>64</u> <u>54</u>	70 72 <u>66</u> <u>61</u>	67 70 68 66	<b>82</b> 79 78 77	<b>84</b> 80 79 76	$     \underline{67}     \underline{64}     \underline{65}     \underline{63}   $	71 69 71 69	<b>89.2</b> 87.5 87.0
Russian Czech French German	<b>98</b> 97 <b>98</b> 89	<b>97</b> 95 93 96 94	93 <b>100</b> 94 96 93	93 99 96 97 88	96 95 92 94 <b>97</b>	<b>88</b> <b>88</b> 87 85 81	87 90 88 <b>89</b> 86	83 90 88 86 78	$     \begin{array}{r}         \underline{66} \\         \underline{68} \\         \underline{64} \\         \underline{54} \\         \underline{59}     \end{array} $	$     \frac{70}{72}     \frac{66}{61}     \underline{62} $	67 70 68 66 <u>57</u>	<b>82</b> 79 78 77 78	<b>84</b> 80 79 76 78	$     \begin{array}{r} \underline{67} \\ \underline{64} \\ \underline{65} \\ \underline{63} \\ \underline{67} \end{array} $	71 69 71 69 68	<b>89.2</b> 87.5 87.0 85.2
Russian Czech French German Arabic	<b>98</b> 97 <b>98</b> 89 85	<b>97</b> 95 93 96 94 86	93 <b>100</b> 94 96 93 88	93 99 96 97 88 85	96 95 92 94 <b>97</b> 89	<b>88</b> <b>88</b> 87 85 81 71	87 90 88 <b>89</b> 86 70	83 90 88 86 78 65	$     \begin{array}{r} \underline{66} \\ \underline{68} \\ \underline{64} \\ \underline{54} \\ \underline{59} \\ \underline{63} \end{array} $	$     \frac{70}{72} \\     \frac{66}{61} \\     \frac{62}{66} \\     \frac{66}{66} $	67 70 68 66 <u>57</u> <u>59</u>	<b>82</b> 79 78 77 78 74	<b>84</b> 80 79 76 78 79		71 69 71 69 68 <u>65</u>	<b>89.2</b> 87.5 87.0 85.2 80.3
Russian Czech French German Arabic Mandarin	<b>98</b> 97 <b>98</b> 89 85 85	<b>97</b> 95 93 96 94 86 85	93 <b>100</b> 94 96 93 88 88	93 99 96 97 88 85 85	96 95 92 94 <b>97</b> 89 86	<b>88</b> <b>88</b> 87 85 81 71 82	87 90 88 <b>89</b> 86 70 87	83 90 88 86 78 65 89	$ \begin{array}{r} \underline{66} \\ \underline{68} \\ \underline{64} \\ \underline{54} \\ \underline{59} \\ \underline{63} \\ 80 \end{array} $	$     \frac{70}{72}     \frac{66}{61}     \frac{61}{62}     \frac{66}{66}     78 $	67 70 68 66 <u>57</u> <u>59</u> 77	<b>82</b> 79 78 77 78 74 74	<b>84</b> 80 79 76 78 79 80	67 64 65 63 67 66 <b>91</b>	71 69 71 69 68 <u>65</u> <b>86</b>	<b>89.2</b> 87.5 87.0 85.2 80.3 84.6

Projecting
 non SOV to strict SOV.

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud	
English	96	95	93	94	94	86	98	98	96	96	96	93	93	95	96	98	
Russian	95	94	86	93	95	90	94	96	95	97	99	94	94	93	96	96	
Czech	94	94	83	95	100	92	92	93	94	94	98	95	88	89	92	94	
French	94	91	84	92	95	90	92	97	95	95	99	94	87	93	96	96	
German	94	87	95	94	88	82	90	95	94	92	93	91	90	89	95	96	
Arabic	90	96	76	85	84	90	85	87	86	85	84	87	84	85	89	85	
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87	
Turkish	67	61	61	65	71	62	64	69	75	77	71	73	73	68	68	74	
Korean	<u>51</u>	53	63	<u>53</u>	61	52	51	54	59	65	<u>59</u>	59	57	55	54	61	
Japanese	<u>55</u>	<u>56</u>	<u>41</u>	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	55	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	54	
	pl	pt	ru	ru	SV	eu	hi	tr	ja	ja	ko	vi	th	zh	zh		
	pud	pud	pud	syntag	pud	bdt	pud	pud	gsd	pud	pud	vtb	pud	gsd	pud	mean	
English	95	97	93	93	96	88	87	83	66	<u>70</u>	67	82	84	67	71	88.9	Project
Russian	98	95	100	99	95	88	90	90	68	72	70	79	80	64	69	<del>89.2</del>	non SC
Czech	97	93	94	96	92	87	88	88	64	<u>66</u>	68	78	79	65	71	87.5	strict S
French	98	96	96	97	94	85	89	86	54	61	66	77	76	63	69	87.0	
German	89	94	93	88	97	81	86	78	59	62	57	78	78	67	68	85.2	
Arabic	85	86	88	85	89	71	70	65	63	66	59	74	79	66	65	80.3	
Mandarin	85	85	86	85	86	82	87	89	80	78	77	74	80	91	86	84.6	
Turkish	76	73	71	71	69	79	83	94	82	83	88	<u>63</u>	<u>68</u>	72	71	72.3	
								10000 Day			122		(1	(7	"	(21	
Korean	66	58	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u> </u>	61	67	<u>66</u>	63.1	Projecting

ing DV to OV.

OV. 15

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud	
English	96	95	93	94	94	86	98	98	96	96	96	93	93	95	96	98	
Russian	95	94	86	93	95	90	94	96	95	97	99	94	94	93	96	96	
Czech	94	94	83	95	100	92	92	93	94	94	98	95	88	89	92	94	
French	94	91	84	92	95	90	92	97	95	95	99	94	87	93	96	96	
German	94	87	95	94	88	82	90	95	94	92	93	91	90	89	95	96	Only
Arabic	90	96	76	85	84	90	85	87	86	85	84	87	84	85	89	85	Mandarin
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87	never
Turkish	67	61	61	65	71	<u>62</u>	64	69	75	77	71	73	73	68	68	74	loses to the
Korean	51	53	63	<u>53</u>	61	<u>52</u>	51	54	<u>59</u>	65	<u>59</u>	59	57	55	54	61	majority
Japanese	55	<u>56</u>	<u>41</u>	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	55	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	54	baseline.
	pl	pt	ru	ru	SV	eu	hi	tr	ja	ja	ko	vi	th	zh	zh	maan	
	pud	pud	pud	syntag	pud	bdt	pud	pud	gsd	pud	pud	vtb	pud	gsd	pud	mean	
English	95	97	93	93	96	88	87	83	66	70	67	82	84	67	71	88.9	Projecting
Russian	98	95	100	99	95	88	90	90	68	72	70	79	80	64	69	89.2	non SOV to
Czech	97	93	94	96	92	87	88	88	64	66	68	78	79	65	71	87.5	strict SOV.
French	98	96	96	97	94	85	89	86	54	61	66	77	76	63	69	87.0	
German	89	94	93	88	97	81	86	78	59	62	57	78	78	67	68	85.2	
				1211-112		71	70	65	63	66	59	74	79	66	65	80.3	
Arabic	85	86	88	85	89	71	10	05	05								
Arabic Mandarin	85 85	86 85	88 86	85 85	89 86	82	87	89	80	78	77	74	80	91	86	84.6	
									-		77 88	74 <u>63</u>	80 <u>68</u>	<b>91</b> 72	<b>86</b> 71	84.6 72.3	
Mandarin	85	85	86	85	86	82	87	89	80	78							Projecting

Projecting SVO to SVO; note that Arabic is not that good across the board.

	ar padt	ga idt	af booms	de pud	cs pud	cy ccg	en ewt	en pud	es pud	fi pud	fr pud	he hdt	hy arm	id pud	is pud	it pud
	-		1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	<u> </u>	-			-			•			-		
English	96	95	93	94	94	86	98	98	96	96	96	93	93	95	96	98
Russian	95	94	86	93	95	90	94	96	95	97	99	94	94	93	96	96
Czech	94	94	83	95	100	92	92	93	94	94	98	95	88	89	92	94
French	94	91	84	92	95	90	92	97	95	95	99	94	87	93	96	96
German	94	87	95	94	88	82	90	95	94	92	93	91	90	89	95	96
Arabic	90	96	76	85	84	90	85	87	86	85	84	87	84	85	89	85
Mandarin	86	84	85	87	87	85	85	86	86	89	87	87	81	83	86	87
Turkish	67	61	61	65	71	<u>62</u>	64	69	75	77	71	73	73	68	68	74
Korean	<u>51</u>	53	63	<u>53</u>	61	<u>52</u>	51	54	<u>59</u>	65	59	59	57	55	54	61
Japanese	<u>55</u>	56	<u>41</u>	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	54	55	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>
	pl	pt	ru	ru	SV	eu	hi	tr	ja	ja	ko	vi	th	zh	zh	
	pl pud	pt pud	ru pud	ru syntag	sv pud	eu bdt	hi pud	tr pud	ja gsd	ja pud	ko pud	vi vtb	th pud	zh gsd	zh pud	mean
English		-							•	•						<b>mean</b> 88.9
English Russian	pud	pud	pud	syntag	pud	bdt	pud	pud	gsd	pud	pud	vtb	pud	gsd	pud	
and the second se	pud 95	pud 97	<b>pud</b> 93	syntag 93	<b>pud</b> 96	bdt 88	<b>pud</b> 87	<b>pud</b> 83	<b>gsd</b>	<b>pud</b>	<b>pud</b> 67	vtb 82	pud 84	<b>gsd</b>	<b>pud</b> 71	88.9
Russian	<b>pud</b> 95 <b>98</b>	<b>pud</b> 97 95	<b>pud</b> 93 <b>100</b>	<b>syntag</b> 93 <b>99</b>	<b>pud</b> 96 95	bdt 88 88	<b>pud</b> 87 90	<b>pud</b> 83 90	<b>gsd</b> 66 68	<b>pud</b> <u> 70</u> 72	<b>pud</b> 67 70	<b>vtb</b> <b>82</b> 79	<b>pud</b> <b>84</b> 80	<b>gsd</b> <u>67</u> <u>64</u>	<b>pud</b> 71 69	88.9 <b>89.2</b>
Russian Czech	<b>pud</b> 95 <b>98</b> 97	<b>pud</b> 97 95 93	<b>pud</b> 93 <b>100</b> 94	<b>syntag</b> 93 <b>99</b> 96	<b>pud</b> 96 95 92	bdt 88 88 87	<b>pud</b> 87 90 88	<b>pud</b> 83 90 88	<b>gsd</b> 66 68 64 54	<b>pud</b> <u>70</u> 72 <u>66</u>	<b>pud</b> 67 70 68	<b>vtb</b> <b>82</b> 79 78	<b>pud</b> <b>84</b> 80 79	gsd 67 64 65	<b>pud</b> 71 69 71	88.9 <b>89.2</b> 87.5
Russian Czech French	<b>pud</b> 95 <b>98</b> 97 <b>98</b>	<b>pud</b> 97 95 93 96	<b>pud</b> 93 <b>100</b> 94 96	<b>syntag</b> 93 <b>99</b> 96 97	<b>pud</b> 96 95 92 94	bdt 88 88 87 85	<b>pud</b> 87 90 88 <b>89</b>	<b>pud</b> 83 90 88 86	gsd 66 68 64	pud           70           72           66           61	<b>pud</b> 67 70 68 66	<b>vtb</b> <b>82</b> 79 78 77	<b>pud</b> <b>84</b> 80 79 76	gsd 67 64 65 63	<b>pud</b> 71 69 71 69	88.9 <b>89.2</b> 87.5 87.0
Russian Czech French German	<b>pud</b> 95 <b>98</b> 97 <b>98</b> 89	<b>pud</b> 97 95 93 96 94	<b>pud</b> 93 <b>100</b> 94 96 93	<b>syntag</b> 93 <b>99</b> 96 97 88	<b>pud</b> 96 95 92 94 <b>97</b>	bdt 88 88 87 85 81	<b>pud</b> 87 90 88 <b>89</b> 86	<b>pud</b> 83 90 88 86 78	<b>gsd</b> <u>66</u> <u>68</u> <u>64</u> <u>54</u> <u>59</u>	pud           70           72           66           61           62	<b>pud</b> 67 70 68 66 <u>57</u>	<b>vtb</b> <b>82</b> 79 78 77 78	<b>pud</b> <b>84</b> 80 79 76 78	gsd 67 64 65 63 67	<b>pud</b> 71 69 71 69 68	88.9 <b>89.2</b> 87.5 87.0 85.2
Russian Czech French German Arabic	<b>pud</b> 95 <b>98</b> 97 <b>98</b> 89 85	<b>pud</b> 97 95 93 96 94 86	<b>pud</b> 93 <b>100</b> 94 96 93 88	<b>syntag</b> 93 <b>99</b> 96 97 88 88 85	<b>pud</b> 96 95 92 94 <b>97</b> 89	bdt 88 87 85 81 71	<b>pud</b> 87 90 88 <b>89</b> 86 70	<b>pud</b> 83 90 88 86 78 65	<b>gsd</b> <u>66</u> <u>68</u> <u>64</u> <u>54</u> <u>59</u> <u>63</u>	pud           70           72           66           61           62           66	<b>pud</b> 67 70 68 66 <u>57</u> <u>59</u>	<b>vtb</b> <b>82</b> 79 78 77 78 78 74	<b>pud</b> <b>84</b> 80 79 76 78 79	<b>gsd</b> 67 64 65 63 67 66	<b>pud</b> 71 69 71 69 68 65	88.9 <b>89.2</b> 87.5 87.0 85.2 80.3
Russian Czech French German Arabic Mandarin	<b>pud</b> 95 <b>98</b> 97 <b>98</b> 89 85 85	<b>97</b> 95 93 96 94 86 85	<b>pud</b> 93 <b>100</b> 94 96 93 88 88 86	<b>syntag</b> 93 <b>99</b> 96 97 88 85 85	<b>pud</b> 96 95 92 94 <b>97</b> 89 86	bdt 88 87 85 81 71 82	<b>pud</b> 87 90 88 <b>89</b> 86 70 87	<b>pud</b> <ul> <li>83</li> <li>90</li> <li>88</li> <li>86</li> <li>78</li> <li>65</li> <li>89</li> </ul>	<b>gsd</b> <u>66</u> <u>68</u> <u>64</u> <u>54</u> <u>59</u> <u>63</u> <u>80</u>	70           72           66           61           62           66           78	<b>pud</b> 67 70 68 66 <u>57</u> <u>59</u> 77	<b>vtb</b> <b>82</b> 79 78 77 78 74 74	<b>pud</b> <b>84</b> 80 79 76 78 79 80	gsd 67 64 65 63 67 66 91	<b>pud</b> 71 69 71 69 68 65 <b>86</b>	88.9 <b>89.2</b> 87.5 87.0 85.2 80.3 84.6

#### Some analysis

Models trained on European languages seem to overemphasize the presence and position of complementisers.

More variety in the source language (Russian vs. English, Mandarin vs. the rest) helps.

## Take aways

- Subordinate-clause detection is an easy task with a long tail of hard cases even in the language-internal setting, which puts an upper limit on zero-shot performance.
- Zero-shot performance is dominated by the word-order typology, especially the SVO vs. SOV divide.
- Future work: check if these hard cases noticeably impact performance on NLU tasks.

## Take aways

- Subordinate-clause detection is an easy task with a long tail of hard cases even in the language-internal setting, which puts an upper limit on zero-shot performance.
- Zero-shot performance is dominated by the word-order typology, especially the SVO vs. SOV divide.
- Future work: check if these hard cases noticeably impact performance on NLU tasks.

