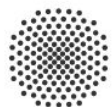


Word-order typology in Multilingual BERT: A case study in subordinate-clause detection

Dmitry Nikolaev, Sebastian Padó

SIGTYP 2022



University of Stuttgart
Institute for Natural Language Processing



Why subordinate-clause detection

The distinction between main and subordinate clauses is arguably universal *conceptually*: subordinate clauses (prototypically) encode *non-asserted information*.[†]

Crucial for claim, stance, and factivity detection.

[†] Cristofaro, S. (2003). Subordination. Oxford: Oxford University Press.

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

→ 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	Vietnamese	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	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	Russian
Accuracy	96.9	97.1	97.4	97.9	98.2	98.8	99	99.1	99.6	99.7

Single-language results

Language	Mandarin	Vietnamese	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	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	Russian
Accuracy	96.9	97.1	97.4	97.9	98.2	98.8	99	99.1	99.6	99.7

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	Vietnamese	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	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	Russian
Accuracy	96.9	97.1	97.4	97.9	98.2	98.8	99	99.1	99.6	99.7

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

Single-language results

Errors mostly due to
annotational discrepancies
between the train and test set.

Language	Mandarin	Vietnamese	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	Spanish	Irish	English	Hebrew	Afrikaans	French	Japanese	Czech	Russian
Accuracy	96.9	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.

Cross-lingual results

Target

Source

	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	<u>51</u>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	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>	<u>55</u>	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>

	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	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	98	95	100	99	95	88	90	90	<u>68</u>	72	70	79	80	<u>64</u>	69	89.2
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	78	79	<u>65</u>	71	87.5
French	98	96	96	97	94	85	89	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	97	81	86	78	<u>59</u>	<u>62</u>	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	<u>66</u>	<u>59</u>	74	79	<u>66</u>	<u>65</u>	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
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u>52</u>	<u>61</u>	<u>67</u>	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	99	98	95	54	<u>70</u>	72	<u>66</u>	60.3

Cross-lingual results

	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	<u>51</u>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	61
Japanese	<u>55</u>	<u>56</u>	41	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	<u>55</u>	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>

	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	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	98	95	100	99	95	88	90	90	<u>68</u>	72	70	79	80	<u>64</u>	69	89.2
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	78	79	<u>65</u>	71	87.5
French	98	96	96	97	94	85	89	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	97	81	86	78	<u>59</u>	<u>62</u>	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	<u>66</u>	<u>59</u>	74	79	<u>66</u>	<u>65</u>	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
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u>52</u>	<u>61</u>	<u>67</u>	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	99	98	95	54	<u>70</u>	72	<u>66</u>	60.3

Russian as the source language has the best overall perf.

Cross-lingual results

	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	<u>51</u>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	61
Japanese	<u>55</u>	<u>56</u>	41	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	<u>55</u>	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>

	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	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	98	95	100	99	95	88	90	90	<u>68</u>	72	70	79	80	<u>64</u>	<u>69</u>	89.2
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	78	<u>79</u>	<u>65</u>	71	87.5
French	98	96	96	97	94	85	89	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	97	81	86	78	<u>59</u>	<u>62</u>	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	<u>66</u>	<u>59</u>	74	79	<u>66</u>	<u>65</u>	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
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u>52</u>	<u>61</u>	<u>67</u>	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	99	98	95	54	<u>70</u>	72	<u>66</u>	60.3

Projecting non SOV to strict SOV.

Cross-lingual results

	ar	ga	af	de	cs	cy	en	en	es	fi	fr	he	hy	id	is	it
	padt	idt	booms	pud	pud	ccg	ewt	pud	pud	pud	pud	hdt	arm	pud	pud	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	<u>51</u>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	61
Japanese	<u>55</u>	<u>56</u>	41	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	<u>55</u>	<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	mean
	pud	pud	pud	syntag	pud	bdt	pud	pud	gsd	pud	pud	vtb	pud	gsd	pud	
English	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	98	95	100	99	95	88	90	90	<u>68</u>	72	70	79	80	<u>64</u>	<u>69</u>	89.2
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	78	<u>79</u>	<u>65</u>	71	87.5
French	98	96	96	97	94	85	89	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	97	81	86	78	<u>59</u>	<u>62</u>	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	<u>66</u>	<u>59</u>	74	79	<u>66</u>	<u>65</u>	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
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u>52</u>	<u>61</u>	67	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	99	98	95	54	<u>70</u>	72	<u>66</u>	60.3

Projecting non SOV to strict SOV.

Projecting SOV to SOV. 15

Cross-lingual results

	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	<u>51</u>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	61
Japanese	<u>55</u>	<u>56</u>	41	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	<u>55</u>	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>

Only Mandarin never loses to the majority baseline.

	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	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	98	95	100	99	95	88	90	90	<u>68</u>	72	70	79	80	<u>64</u>	<u>69</u>	89.2
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	<u>78</u>	<u>79</u>	<u>65</u>	71	87.5
French	98	96	96	97	94	85	89	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	97	81	86	78	<u>59</u>	<u>62</u>	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	<u>66</u>	<u>59</u>	74	79	<u>66</u>	<u>65</u>	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
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u>52</u>	<u>61</u>	67	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	99	98	95	54	<u>70</u>	72	<u>66</u>	60.3

Projecting non SOV to strict SOV.

Projecting SOV to SOV.

Cross-lingual results

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
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>	<u>53</u>	63	<u>53</u>	61	<u>52</u>	<u>51</u>	<u>54</u>	<u>59</u>	65	<u>59</u>	59	<u>57</u>	<u>55</u>	<u>54</u>	61
Japanese	<u>55</u>	<u>56</u>	41	<u>39</u>	<u>52</u>	<u>63</u>	<u>51</u>	<u>52</u>	<u>52</u>	<u>54</u>	<u>55</u>	<u>54</u>	<u>58</u>	<u>53</u>	<u>51</u>	<u>54</u>

	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	95	97	93	93	96	88	87	83	<u>66</u>	<u>70</u>	67	82	84	<u>67</u>	71	88.9
Russian	98	95	100	99	95	88	90	90	<u>68</u>	72	70	79	80	<u>64</u>	69	89.2
Czech	97	93	94	96	92	87	88	88	<u>64</u>	<u>66</u>	68	78	79	<u>65</u>	71	87.5
French	98	96	96	97	94	85	89	86	<u>54</u>	<u>61</u>	66	77	76	<u>63</u>	69	87.0
German	89	94	93	88	97	81	86	78	<u>59</u>	<u>62</u>	<u>57</u>	78	78	<u>67</u>	68	85.2
Arabic	85	86	88	85	89	71	70	65	<u>63</u>	<u>66</u>	<u>59</u>	74	79	<u>66</u>	<u>65</u>	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
Korean	66	<u>58</u>	<u>59</u>	<u>59</u>	<u>53</u>	74	76	94	87	88	88	<u>52</u>	<u>61</u>	<u>67</u>	<u>66</u>	63.1
Japanese	<u>54</u>	<u>52</u>	<u>55</u>	<u>55</u>	<u>50</u>	<u>57</u>	63	88	99	98	95	54	<u>70</u>	72	<u>66</u>	60.3

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.

Thank you!