

Automatic Generation of a Graded Reader in Old Church Slavonic

Iglika Nikolova-Stoupak Eva Schaeffer-Lacroix

Gaël Lejeune Aliona Shestakova-Stukun



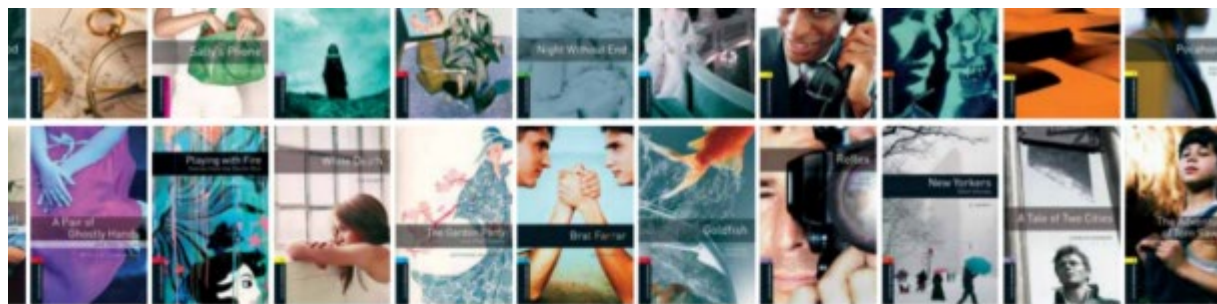
CONTENT

- Context and Motivation
- Data
- Experiments
 - *Automatic Abridgement*
 - *Qualitative Evaluation*
 - *Quantitative Textual Analysis*
- Conclusions

CONTEXT

GRADED READERS

- Typically grouped by **proficiency level** (e.g. CEFR's A1, A2, ...) and/or number of words
- Often **adapted versions** of famous literary texts
- Graded readers offer significant **benefits**, especially in terms of **motivation** and **vocabulary acquisition** (Godwin-Jones, 2018 ; Dickinson, 2017 ; Hill, 2013)
- Graded readers in **classical ('dead') languages** exist (e.g. *Lingua latina per se illustrata*; Athenaze)



Oxford Bookworms Library

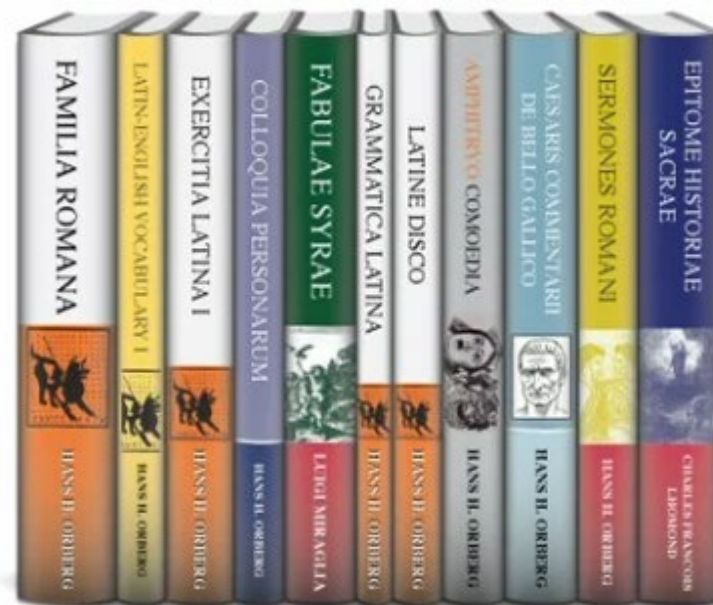
The best and perfect book to learn English



**Oxford
Graded
Readers**



*Lectures CLE
en "français facile"*



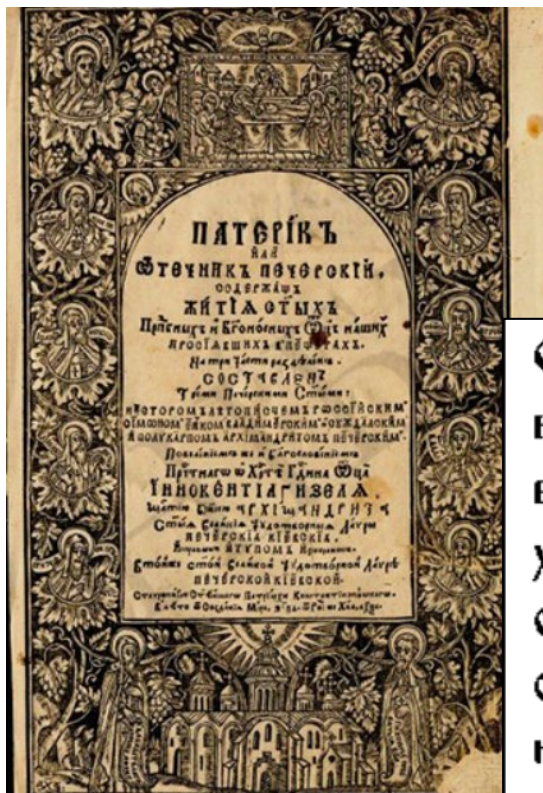
AI AND GRADED READERS?

- the **challenge**: very few graded readers in the majority of languages
- AI could **increase the number** of works and **reduce the time and effort** of professionals
- **ChatGPT**: strong potential with **summarisation** and **simplification**; current problems with **creative writing** and **long contexts** (Achiam et al., 2024 ; Bang et al., 2023 ; Nikolova-Stoupak et al., 2024; Nelson et al., 2023)

THE OLD CHURCH SLAVONIC (OCS) LANGUAGE

- the language of the first Slavic manuscripts (10th-11th centuries)
- large geographic territory, dialects
- written in the Cyrillic (earlier, the Glagolitic) alphabet
- manuscripts: Biblical translations, Saints' lives, prayers and sermons
- developed into distinct 'Church Slavonic' languages

THE OLD CHURCH SLAVONIC (OCS) LANGUAGE



Ѧ	Ѣ	Ѥ	Ѧ	Ѣ	Ѧ
A	B	C	D	E	F
Ѧ	Ѣ	Ѥ	Ѧ	Ѣ	Ѧ
G	H	I	J	K	L
Ѧ	Ѣ	Ѥ	Ѧ	Ѣ	Ѧ
M	N	O	P	R	
Ѧ	Ѣ	Ѥ	Ѧ	Ѣ	Ѧ
					Z

ОТЪЧЕ НАШЪ, НЖЕ КСИ НА НЕБЕСЬХЪ: ДА СВАТИТЪ СА
 НМА ТВОЕ. ДА ПРИДЕТЪ ЦѢСАРЬСТВІЕ ТВОЕ. ДА
 БЖДЕТЪ ВОЛА ТВОЯ, ТАКЪ НА НЕБЕСИ І НА ЗЕМЛИ.
 ХЛѢБЪ НАШЪ НАСТОЯЩАКЪ ДНѢ ДАЖДЪ НАМЪ ДНЬ
 СЪ. І ОТЪПОУСТИ НАМЪ ДЛЪГЫ НАША, ТАКЪ Н МЫ
 ОТЪПОУЩАЕМЪ ДЛЪЖНИКОМЪ НАШИМЪ. І НЕ ВЪВЕДИ
 НАСЪ ВЪ НАПАСТЬ НЪ ИЗБАВИ НЫ ОТЪ НЕПРИАЗНИ.



DATA

GOLD STANDARD

- Classical language readers

<i>language</i>	<i>text</i>	<i>author, year</i>
Latin	<i>Lingua latina per se illustrata</i>	Ørberg, 2011
	<i>Fabulae Faciles</i>	Ritchie, 1903
Greek	<i>Athenaze</i>	Balme & Lawall, 2003
	<i>Logos</i>	Martínez, 2023
Hebrew	<i>Graded Reader of Biblical Hebrew</i>	Van Pelt & Pratico, 2006
	<i>Biblical Hebrew Easy Stories</i>	Aleph with Beth, 2020

ORIGINAL OCS TEXTS

- Chapter 1 of the Biblical book of 'Genesis'¹
- 'The Legend of Saint George and the Dragon'²



¹Spevák, Tomáš. *Genesis — Chapter 1, Reconstructed Old Church Slavonic Text*. Old Slavic Library.

²Rystencko, Alexander V. *Legenda o sv. Georgii i drakone v vizantiiskoi i slaviano-russkoi literaturakh*. Odessa: "Ekon." Printing House, 1909.

AUTOMATIC ABRIDGEMENT

12

MODEL

- ChatGPT (GPT-5; official chatbot interface)

CEFR LEVELS

- A1 ('Genesis')
- B1 ('Saint George and the Dragon')

ONE-SHOT PROMPTING

- Provides the model of an **example of output**
- Shown to cause the output to resemble the provided example (Nikolova-Stoupak et al, 2024)
- Two chapters of *Lingua latina per se illustrata* used

PROMPT

Please write an adapted version of the following text in Old Church Slavonic, which is suitable for a language learner:

{text}

The difficulty should be similar to that of the following Latin text:

{one-shot example}

QUALITATIVE EVALUATION

QUALITATIVE EVALUATION

- In-depth analysis
- Manual corrections applied

Dimension /Question	Definition
Understandability	
Is the text generally understandable (in terms of storyline, vocabulary, and syntax)?	The text is easy to understand. It is free from overly complex words and sentence constructions.
Correctness	
Does the text exhibit correct vocabulary, grammar, punctuation and structure?	There are no non-existent or irrelevant words, grammatical mistakes, mistakes in punctuation or non-optimal paragraph/sentence structure.
Consistency	
Is the text consistent in terms of register, writing style and the information presented?	The text maintains the same genre/level of formality/verb tenses/character traits and knowledge. There are no contradictions in the presented information. The text is consistent in terms of dialect.
Textual Coherence	
Are textual transitions present and natural? Is anaphora clearly resolvable? Is the text free of needless repetition?	The text flows fluently between sentences (paragraphs). There are no unclear anaphora referents. Information (incl. separate words) is not overly repeated.
Aesthetic Appeal	
Does the text have appropriate length/pacing/style/register? Is the text engaging?	The text feels neither too long nor too short. Appropriate textual length is accorded to plot segments. Formality levels are natural.

QUALITATIVE EVALUATION

Strengths

- A1 text perceptibly simpler
- Short sentences, parallel structures (A1)
- Longer sentences, dialogue (B1)

QUALITATIVE EVALUATION

Stylistic Issues

- Lack of action (B1)
- Instances of high complexity (e.g. verb-less sentences)
- Letters not typical to the language (я, ё, й)
- Pairs of interchangeable letters (ou and Ѹ ; з and ꙗ)
- Reflexive particle merged (съберетъся instead of съберетъ са)
- Imperfective forms shortened (живѣше instead of живѣаше)

QUALITATIVE EVALUATION

Errors

- Confusion of the letters ъ and ь
- Use of the letter я
- Interference of modern languages (меншее instead of мьнѹю)
- Wrong cases (земла instead of землѹ; nom. case)
- Spelling mistakes (четвѣртѹи instead of четврьтѹи)
- Word choice (госпожа instead of дѣвица)

‘Genesis’

Original	12 И рече Ёъ: да бжджть свѣтила на небеси, свѣтити <i>земляж</i> . 13 И сътвори Ёъ два свѣтила: свѣтило <i>вѣлико днѣ</i> , и свѣтило <i>меншее</i> нощи, и <i>звѣзды</i> . 14 И виде Ёъ, <i>яко</i> добро. И бѣ вечерь и бѣ <i>оутро</i> , днь <i>четвѣртый</i> .
----------	--

Corrected	12 И рече Ёъ: да бжджть свѣтила на небеси, свѣтити <i>земляж</i> . 13 И сътвори Ёъ два свѣтила: свѣтило <i>велико дни</i> , и свѣтило <i>мынѣе</i> нощи, и <i>звѣзды</i> . 14 И виде Ёъ, <i>яко</i> добро. И бѣ вечерь и бѣ <i>ѣтро</i> , днь <i>четвертый</i> .
-----------	--

‘Saint George and the Dragon’

Original	Бысть градъ Ласиа, и в <i>нѣмъ</i> царъ именемъ Соломонъ. Близъ града <i>бѣше</i> озеро велико, и в томъ озерѣ <i>живѣше</i> змиище <i>лють</i> .
----------	---

Corrected	Бысть градъ Ласиа, и в <i>нѣмъ</i> царъ именемъ Соломонъ. Близъ града <i>бѣаше</i> озеро велико, и в томъ озерѣ <i>живѣаше</i> змиище <i>люто</i> .
-----------	---

БЪТИЕ, ГЛАВА 1

1 Въ началѣ сътвори Ёѣ небо и землѣ.

2 Земля бѣ пуста и тѣма бѣ на водахъ. Дѣхъ бѣи хождаеше на водахъ.

3 И рече Ёѣ: да бѣдетъ свѣтъ. И бѣ свѣтъ.

4 И виде Ёѣ свѣтъ, ꙗко добръ. И раздѣчи Ёѣ свѣтъ отъ тѣмы.

5 И нарече Ёѣ свѣтъ днь, а тѣмъ нарече нощъ. И бѣ вечеръ и бѣ ътро, днь едѣнъ.

6 И рече Ёѣ: да бѣдетъ тврьдъ посрѣдѣ водъ. И бѣ тврьдъ.

7 И нарече Ёѣ тврьдъ небо. И бѣ вечеръ и бѣ ътро, днь вторѣи.

8 И рече Ёѣ: да съберетъ сѧ вода, и да ꙗвить сѧ сѣша. И бѣ тако.

9 И нарече Ёѣ сѣшъ землѣ, а воды нарече морѣ. И виде Ёѣ, ꙗко добро.

10 И рече Ёѣ: да прорастиъ земля травъ и древо плодоносно. И бѣ тако.

11 И прорастиъ земля травъ и дрѣво. И виде Ёѣ, ꙗко добро. И бѣ вечеръ и бѣ ътро, днь третѣи.

12 И рече Ёѣ: да бѣдѣтъ свѣтила на небеси, свѣтити землѣ.

13 И сътвори Ёѣ два свѣтила: свѣтило велико дни, и свѣтило мѣнѣе нощи, и звѣзды.

14 И виде Ёѣ, ꙗко добро. И бѣ вечеръ и бѣ ътро, днь четвѣртѣи.

15 И рече Ёѣ: да изведѣтъ воды рыбѣ, и да летаѣтъ птица надъ землѣ.

16 И сътвори Ёѣ рыбѣ и птицѣ. И благослови ꙗ Ёѣ: растите сѧ и мѣножите сѧ.

17 И бѣ вечеръ и бѣ ътро, днь пятѣи.

18 И рече Ёѣ: да изведѣтъ земля скоть и звѣрь.

19 И сътвори Ёѣ скоть и звѣрь. И виде Ёѣ, ꙗко добро.

20 И рече Ёѣ: сътворимъ чловѣка по образъ нашъ.

21 И сътвори Ёѣ чловѣка: мѣжа и женѣ.

22 И благослови ꙗ Ёѣ: растите сѧ и плодите сѧ, и наполните землѣ.

23 И виде Ёѣ всѧ, ꙗже сътвори. И се бысть добро зѣло.

24 И бѣ вечеръ и бѣ ътро, днь шѣтѣи.

Чюдо етаго Гевргіа

Бысть градъ Ласиа, и в ньмъ царь именемъ Соломонъ. Близъ града бѣаше озеро велико, и в томъ озерѣ живѣаше змище люто.

И сходяше змище на землю, и всадневно пожираше люди. И рекоша граждане къ царю:
— Цаю, аше не далимъ змиишч чловѣка. вси погибнемъ.

И метааху жребии, и по жребию давали дѣти свои змиищу. Наконѣцъ жребии паде на дщерь цареву. Царь плакаше горько и рече:
— Чадѡ моѡ, иди нынѣ къ змиищу.

И обрѣтена бѣ дѣва у озера, плачущиѡ сѡ.

И прииде сѣи Гевргіи, воинъ христіанинъ. И рече къ дѣвѣ:
— Что плачеши, отроковице?

Рече дѣва:

— Здѣ обрѣтаеть мѡ змиище, и погыбну. Бѣжи скоро, добрыи чловѣче, да и ты не умреш.

Рече сѣи Гевргіи:

— Не бои сѡ, дѣво. Поможеть намъ Христѡсъ.

И се змиище изыде изъ воды, велико и страшно. Сѣи же Гевргіи сотвори крѣстное знамение и рече:

— Господи Іисусе Христе, помози ми.

И устреми конѡ своего, и копіемъ порази змиѡ въ главу, и низверже его на землю.

И рече сѣи къ дѣвѣ:

— Възложи подѡсъ твои на змиище и влечи во градъ.

И видѣша людие чюдо велие. Удивиша сѡ и устраиша сѡ. Рече же сѣи Гевргіи:

— Не боите сѡ. Вѣруите во Господа нашего Іисуса Христа, и крестите сѡ, и избавить васъ отъ всякаго зла.

И вѣрова царь и вѣрова людие, и крестивши сѡ. И бѣ радость велика во граде.

QUANTITATIVE TEXTUAL ANALYSIS

PRE-PROCESSING

- If necessary, OCR (tool *Pen-to-Print*)
- Retaining solely of target language text (no prefaces, exercises, glossaries, titles)
- Standardisation of punctuation (e.g. Greek ‘;’ changed to ‘?’)
- Text merged (no line breaks and tabs)
- Multiple volumes merged into one

TEXTUAL CHARACTERISTICS

- Readability-based
- Not exhaustive
- Open-source pipeline UDPipe employed

Model: ☒ UD 2.15 (docs) ☐ UD 2.12 (docs) ☐ UD 2.10 (docs) ☐ UD 2.6 (docs) ☐ PDT-C 1.0 (docs) ☐ EvaLatin (24/20)

latin-perseus-ud-2.15-241121

Actions: ☒ Tag and Lemmatize ☒ Parse

▼ Advanced Options

Input Text Input File

Roma in Italia est. Italia in Europa est. Graecia in Europa est. Italia et Graecia in Europa sunt. Hispania quoque in Europa est. Hispania et Hispania et Italia et Graecia in Europā sunt.

Output Text Show Table Show Trees

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe									
# udpipe_model = latin-perseus-ud-2.15-241121									
# udpipe_model_licence = CC BY-NC-SA									
# newdoc									
# newpar									
# sent_id = 1									
# text = Roma in Italia est.									
1	Roma	Roma	PROPN	n-s---fn-	Case=Nom Gender=Fem Number=Sing	3	nsubj	_	TokenRange=0:4
2	in	in	ADP	r-----	_	3	case	_	TokenRange=5:7
3	Italia	Italia	PROPN	n-s---fb-	Case=Ab Gender=Fem Number=Sing	0	root	_	TokenRange=8:14
4	est	sum	AUX	v3spia---	Aspect=Imp Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	3	cop	_	SpaceAfter=No TokenRange=15:18
5	.	.	PUNCT	u-----	_	3	punct	_	TokenRange=18:19

TEXTUAL CHARACTERISTICS

Type of textual feature	Selected features
Length-based	average number of letters per word average number of words per sentence
Vocabulary-related	word-based TTR lemma-based TTR
Syntax-related	average number of verbs per sentence average percentage of function words per sentence
Discourse-related	average number of pronouns per sentence punctuation variety

TEXTUAL CHARACTERISTICS

Text	<i>LLPSI</i>	<i>LLPSI</i>	<i>Fabulae Faciles</i>	<i>Athenaze</i>	<i>Athenaze</i>	<i>Logos</i>	<i>Logos</i>	<i>GRBH</i>	<i>GRBH</i>	<i>BHES</i>	“Genesis: 1”	“Saint George and the Dragon”
Level	A1	B1	A1	A1	B1	A1	B1	A1	B1	A1	A1	B1
Language	Latin	Latin	Latin	Greek	Greek	Greek	Greek	Hebrew	Hebrew	Hebrew	OCS	OCS
Avg letters/word	5.23	5.35	5.66	4.78	5.20	4.58	4.97	4.42	4.53	4.37	3.70	4.37
Avg words/s-ce	12.51	14.49	18.68	35.01	25.22	12.53	17.42	13.43	13.70	9.08	7.54	9.00
TTR (words)	0.36	0.42	0.40	0.40	0.52	0.33	0.40	0.50	0.57	0.28	0.42	0.69
TTR (lemmas)	0.21	0.25	0.27	0.30	0.40	0.25	0.30	0.43	0.50	0.20	0.37	0.57
Avg verbs/s-ce	1.43	1.84	3.19	4.08	4.20	1.33	2.29	2.48	2.10	2.46	1.23	1.65
Avg % funct. words/s-ce	27.89	30.38	34.54	33.11	33.11	31.31	33.78	30.86	22.34	32.50	33.51	33.49
Avg pronouns/s-ce	0.44	0.59	0.75	1.46	1.65	0.63	0.84	1.26	1.16	1.29	0.35	0.62
Punctuation variety	3.02	3.23	2.02	6.66	3.19	3.71	3.46	2.83	2.95	2.13	0.97	1.65
Total s-ces	905	1317	366	198	107	458	487	92	98	728	40	29

TEXTUAL CHARACTERISTICS

- ChatGPT output has **lower values** than gold standard for **length-based features**
- **High TTR** for 'Saint George and the Dragon'
- **Low verbs and pronouns** per sentence for 'Genesis'
- **Low punctuation variety** for both texts
- **Trends** associated with A1 vs B1 matched

CONCLUSIONS

- ChatGPT's general **abridgement abilities** demonstrated
- **Differences between levels A1 and B1** match the gold standard
- **Existent notions** of Old Church Slavonic
- Some **mistakes** and **stylistic issues**
- **Interference** with modern languages
- High textual difficulty for level B1
- Professional intervention is a necessity

Thank you for your kind attention!

Do you have questions?

