TEXT-BASED GLOSSARY GENERATION BY CHATGPT FOR FL LEARNERS

Iglika Nikolova-Stoupak

Gaël Lejeune

Eva Schaeffer-Lacroix

SORBONNE UNIVERSITÉ

Sens Texte Informatique Histoire

INNOVATION IN LANGUAGE LEARNING International Conference

Outline

- Background (extensive reading and glosses)
- Corpus (texts with glosses)
- Analysis of the glosses
- Glossary generation by ChatGPT
  - zero-shot
  - one-shot
  - chain-of-thought
- Conclusions and future work

Background: extensive reading

- unintentional acquisition
- vocabulary as it appears in context
- e.g. graded readers

Limitations

- need to read a large number of texts
- slow progress
- possible wrong inferences

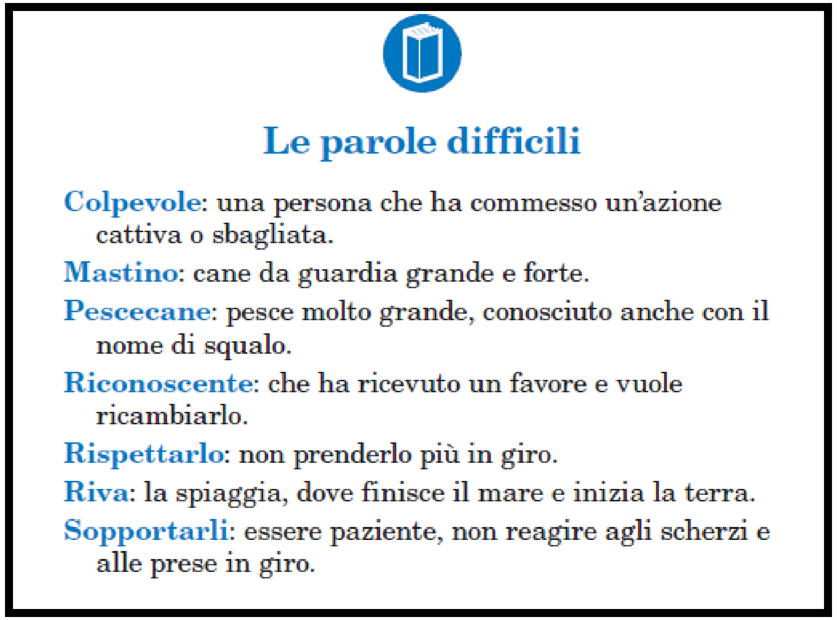
Solutions

- teacher involvement
- in-class environment
- added intentional acquisition

Glosses

- supplementary lexical information about vocabulary items, such as definitions, translations and/or illustrations

- addressed by the field of lexicography
- formats:
  - glossaries
  - footnotes
  - …
- decisions to take:
  - how many words/expressions?
  - which words/expressions? (the most complex? the most useful?)



Le parole difficili

Colpevole: una persona che ha commesso un'azione cattiva o sbagliata.

Mastino: cane da guardia grande e forte.

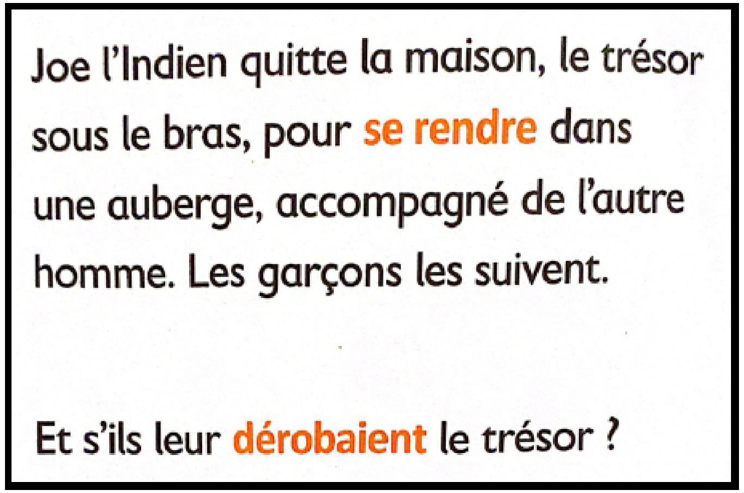
Pescecane: pesce molto grande, conosciuto anche con il nome di squalo.

Riconoscente: che ha ricevuto un favore e vuole ricambiarlo.

Rispettarlo: non prenderlo più in giro.

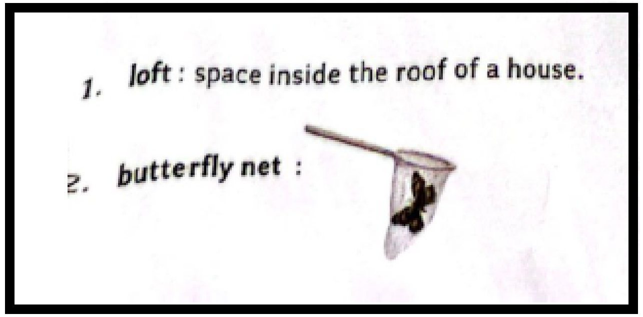
Riva: la spiaggia, dove finisce il mare e inizia la terra.

Sopportarli: essere paziente, non reagire agli scherzi e alle prese in giro.

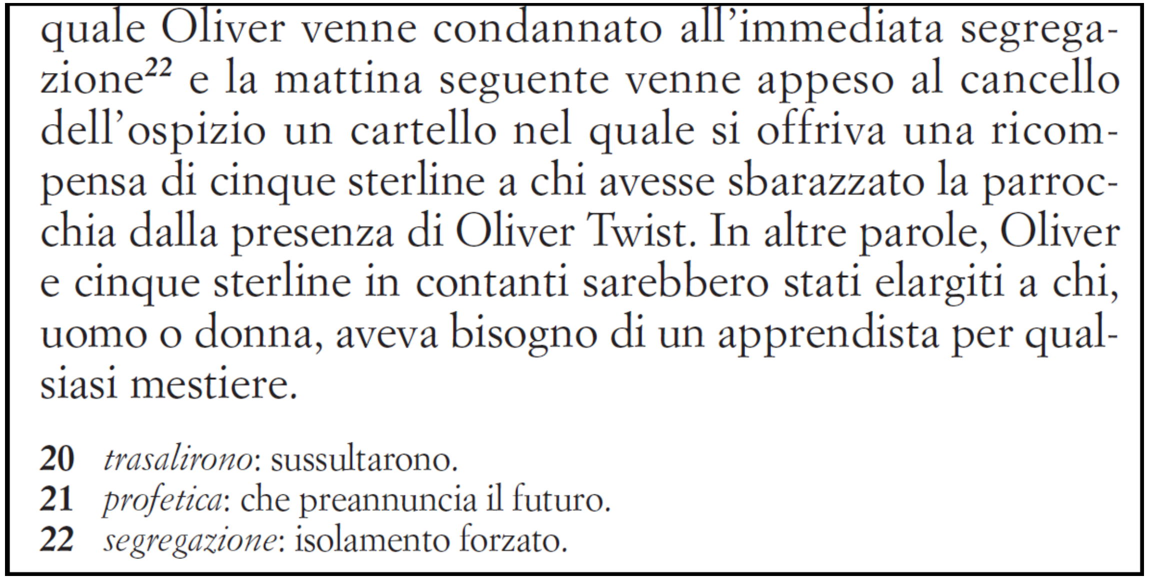


Joe l'Indien quitte la maison, le trésor sous le bras, pour **se rendre** dans une auberge, accompagné de l'autre homme. Les garçons les suivent.

Et s'ils leur **dérobaient** le trésor ?



1. *loft* : space inside the roof of a house.
2. *butterfly net* :

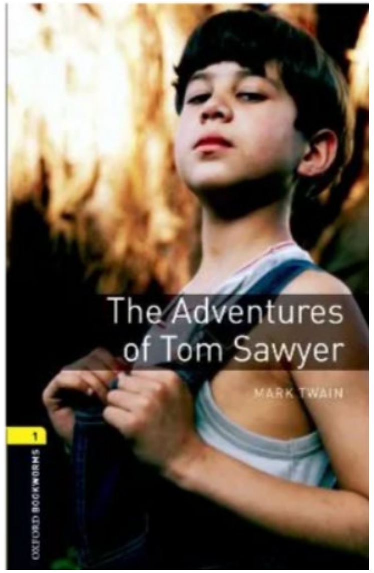


quale Oliver venne condannato all'immediata segregazione[22] e la mattina seguente venne appeso al cancello dell'ospizio un cartello nel quale si offriva una ricompensa di cinque sterline a chi avesse sbarazzato la parrocchia dalla presenza di Oliver Twist. In altre parole, Oliver e cinque sterline in contanti sarebbero stati elargiti a chi, uomo o donna, aveva bisogno di un apprendista per qualsiasi mestiere.

20 *trasalirono*: sussultarono.
21 *profetica*: che preannuncia il futuro.
22 *segregazione*: isolamento forzato.

Corpus Used for Analysis

- 10 texts for language learners or otherwise simplified
- adaptations of works commonly used in academic curricula
- contain glosses
- 4 languages:
    - English
    - Spanish
    - Italian
    - French



Corpus Used for Analysis

title	total words	lang.	author (original)	series	year
A Christmas Carol	6153	🇬🇧	Charles Dickens	Czytamyw oryginale	2004
Don Chisciotte della Mancia	2994	🇮🇹	Miguel de Cervantes	I coriandoli	2005
Les Misérables	9931	🇫🇷	Victor Hugo	Lectures CLE en français facile	2016
Madame Bovary	16529	🇪🇸	Gustave Flaubert	-	2019
Oliver Twist	27071	🇬🇧	Charles Dickens	Oxford Bookworms	2007
Oliver Twist	20750	🇪🇸	Charles Dickens	Kalafate	2016
Romeo y Julieta	16386	🇪🇸	William Sharespeare	Kalafate	2009
The Picture of Dorian Gray	10958	🇬🇧	Oscar Wilde	Oxford Bookworms	2007
Tom Sawyer	16790	🇪🇸	Mark Twain	Kalafate	2005
Tom Sawyer	6284	🇬🇧	Mark Twain	Oxford Bookworms	2000

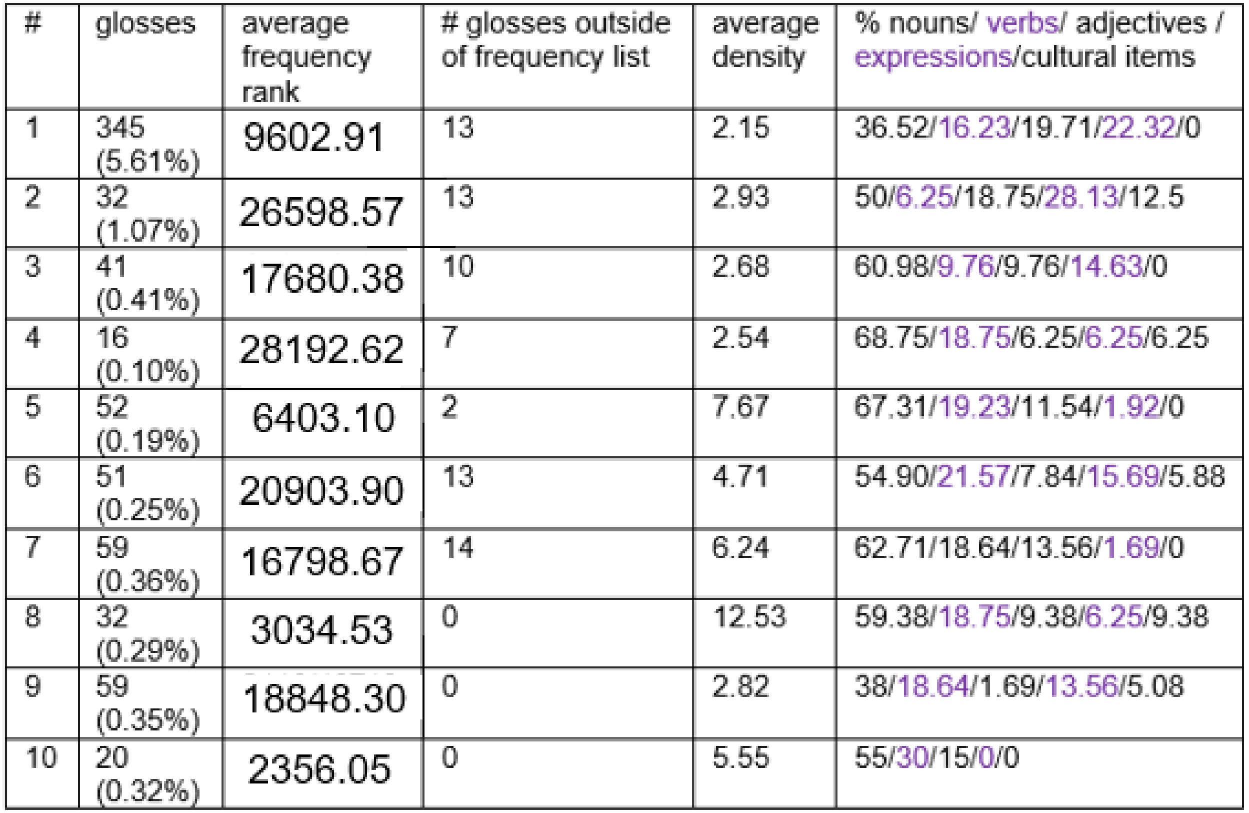
Analysis of the Glosses: methods

qualitative	quantitative
placement	number
content	parts of speech
	density
	frequency ranks

Results: qualitative

placement	content
glossaries at the end of the book (4)	definitions (4)
footnotes (3)	items used in sentence (3)
definitions in margin (3)	translations (1)
	illustrations (3)

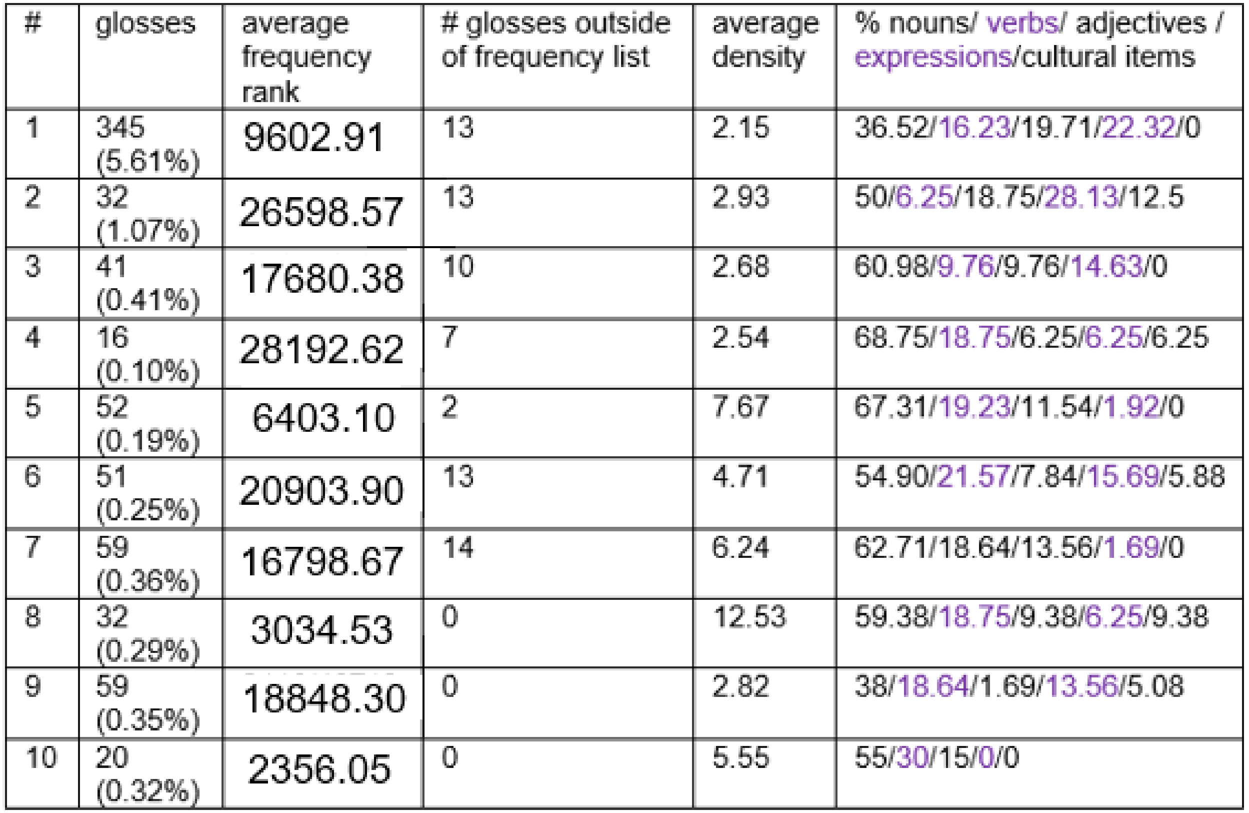
Results: quantitative



#	glosses	average frequency rank	# glosses outside of frequency list	average density	% nouns/ verbs/ adjectives / expressions/cultural items
1	345 (5.61%)	9602.91	13	2.15	36.52/16.23/19.71/22.32/0
2	32 (1.07%)	26598.57	13	2.93	50/6.25/18.75/28.13/12.5
3	41 (0.41%)	17680.38	10	2.68	60.98/9.76/9.76/14.63/0
4	16 (0.10%)	28192.62	7	2.54	68.75/18.75/6.25/6.25/6.25
5	52 (0.19%)	6403.10	2	7.67	67.31/19.23/11.54/1.92/0
6	51 (0.25%)	20903.90	13	4.71	54.90/21.57/7.84/15.69/5.88
7	59 (0.36%)	16798.67	14	6.24	62.71/18.64/13.56/1.69/0
8	32 (0.29%)	3034.53	0	12.53	59.38/18.75/9.38/6.25/9.38
9	59 (0.35%)	18848.30	0	2.82	38/18.64/1.69/13.56/5.08
10	20 (0.32%)	2356.05	0	5.55	55/30/15/0/0

- book 1 is an outlier (345 glosses)
- highest number of nouns (typically above 50%), most commonly followed by verbs
- half of the texts contain “cultural items” (e.g. “Prince Charming”, “the river Thames”)

Results: quantitative



#	glosses	average frequency rank	# glosses outside of frequency list	average density	% nouns/ verbs/ adjectives / expressions/cultural items
1	345 (5.61%)	9602.91	13	2.15	36.52/16.23/19.71/22.32/0
2	32 (1.07%)	26598.57	13	2.93	50/6.25/18.75/28.13/12.5
3	41 (0.41%)	17680.38	10	2.68	60.98/9.76/9.76/14.63/0
4	16 (0.10%)	28192.62	7	2.54	68.75/18.75/6.25/6.25/6.25
5	52 (0.19%)	6403.10	2	7.67	67.31/19.23/11.54/1.92/0
6	51 (0.25%)	20903.90	13	4.71	54.90/21.57/7.84/15.69/5.88
7	59 (0.36%)	16798.67	14	6.24	62.71/18.64/13.56/1.69/0
8	32 (0.29%)	3034.53	0	12.53	59.38/18.75/9.38/6.25/9.38
9	59 (0.35%)	18848.30	0	2.82	38/18.64/1.69/13.56/5.08
10	20 (0.32%)	2356.05	0	5.55	55/30/15/0/0

- English glosses tend to include more common words
- no significant differences by language
- polysemy (e.g. “will” – noun and auxiliary verb) is a problem with frequency lists

Glossary Generation by ChatGPT (GPT-4o): methods

- two scenarios: zero-shot and one-shot
- source text:
  - the first chapter of *Alice in Wonderland*
  - professional translations into Bulgarian and Japanese
  - estimated as CEFR level B1 (based on *Road to Grammar*, *Cathoven* and *Poodll*)
- secondary text:
  - the first two chapters of *The Picture of Dorian Gray* + *glossary*
- additional prompting (chain-of-thought)
  - asking the model to edit the output based on observations from the professional texts

Glossary Generation by ChatGPT (GPT-4o): prompts

Please generate a glossary for the following text. Make it suitable for learners of {English/Japanese/Bulgarian} as a foreign language whose current level is B1.

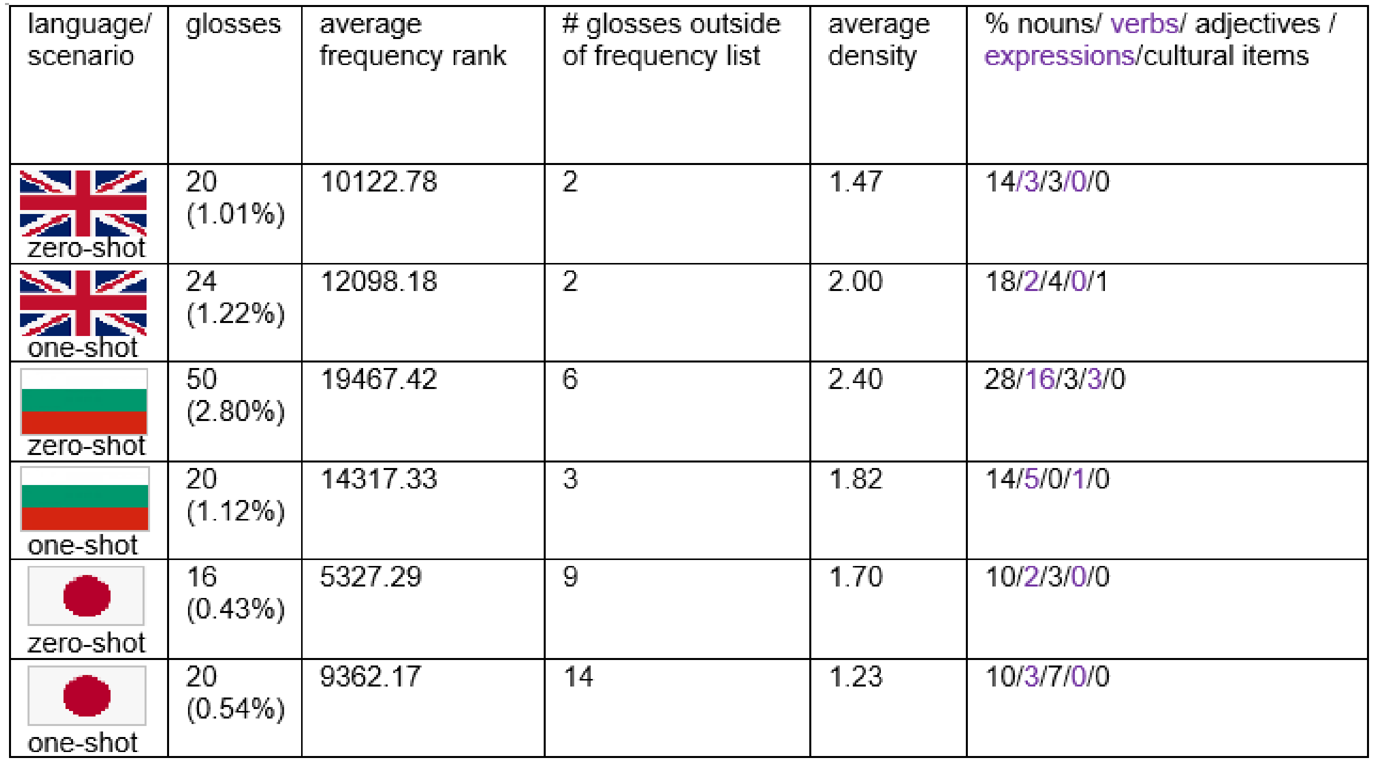
{source text}

(one-shot:)

This is an example of a text followed by a good glossary:

{secondary text + glossary}

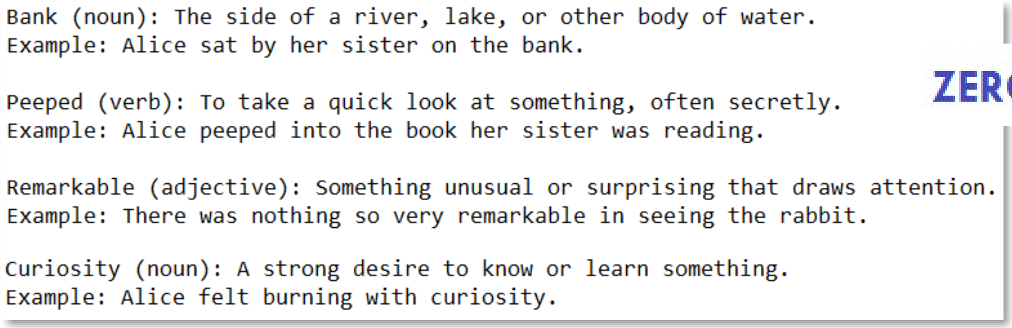
Automatically Generated Glosses



language/ scenario	glosses	average frequency rank	# glosses outside of frequency list	average density	% nouns/ verbs/ adjectives / expressions/cultural items
🇬🇧 zero-shot	20 (1.01%)	10122.78	2	1.47	14/3/3/0/0
🇬🇧 one-shot	24 (1.22%)	12098.18	2	2.00	18/2/4/0/1
🇧🇬 zero-shot	50 (2.80%)	19467.42	6	2.40	28/16/3/3/0
🇧🇬 one-shot	20 (1.12%)	14317.33	3	1.82	14/5/0/1/0
🇯🇵 zero-shot	16 (0.43%)	5327.29	9	1.70	10/2/3/0/0
🇯🇵 one-shot	20 (0.54%)	9362.17	14	1.23	10/3/7/0/0

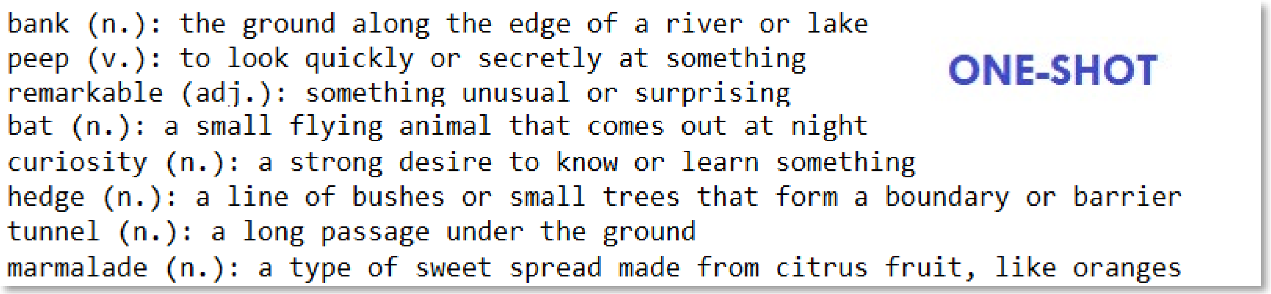
- often round number of items
- mostly definitions
- non-English text – often translations into / definitions in English
- more information in zero-shot scenario (e.g. example sentences)
- average density stable
- POS similar to baseline
- general lack of cultural information

Which Scenario Provided Better Results?



```
Bank (noun): The side of a river, lake, or other body of water.
Example: Alice sat by her sister on the bank.

Peeped (verb): To take a quick look at something, often secretly.
Example: Alice peeped into the book her sister was reading.

Remarkable (adjective): Something unusual or surprising that draws attention.
Example: There was nothing so very remarkable in seeing the rabbit.

Curiosity (noun): A strong desire to know or learn something.
Example: Alice felt burning with curiosity.
```

ZERO-SHOT



```
bank (n.): the ground along the edge of a river or lake
peep (v.): to look quickly or secretly at something
remarkable (adj.): something unusual or surprising
bat (n.): a small flying animal that comes out at night
curiosity (n.): a strong desire to know or learn something
hedge (n.): a line of bushes or small trees that form a boundary or barrier
tunnel (n.): a long passage under the ground
marmalade (n.): a type of sweet spread made from citrus fruit, like oranges
```

ONE-SHOT

One-Shot

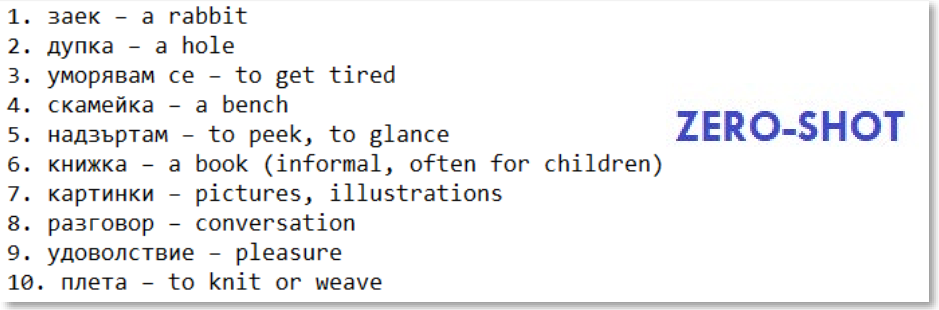
√ higher average density

√ higher variance between minimal and maximal density

√ no example sentences

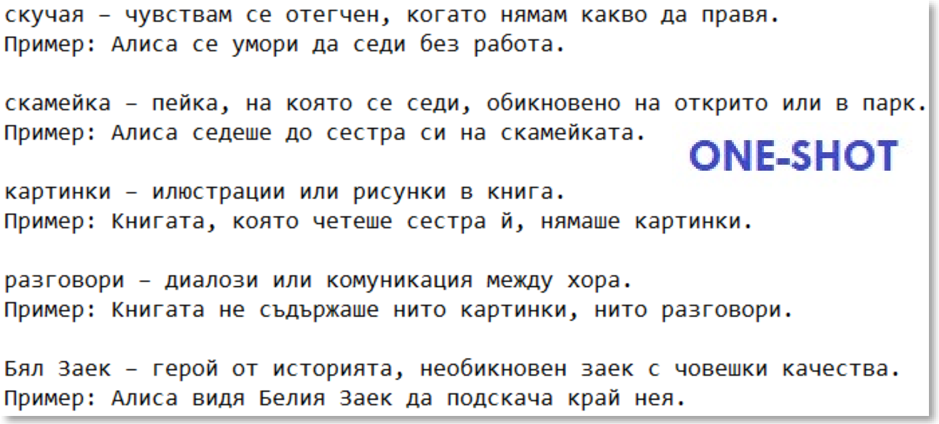
√ a cultural item

Which Scenario Provided Better Results?



ZERO-SHOT

```
1. заек – a rabbit
2. дупка – a hole
3. уморявам се – to get tired
4. скамейка – a bench
5. надзъртам – to peek, to glance
6. книжка – a book (informal, often for children)
7. картинки – pictures, illustrations
8. разговор – conversation
9. удоволствие – pleasure
10. плета – to knit or weave
```



ONE-SHOT

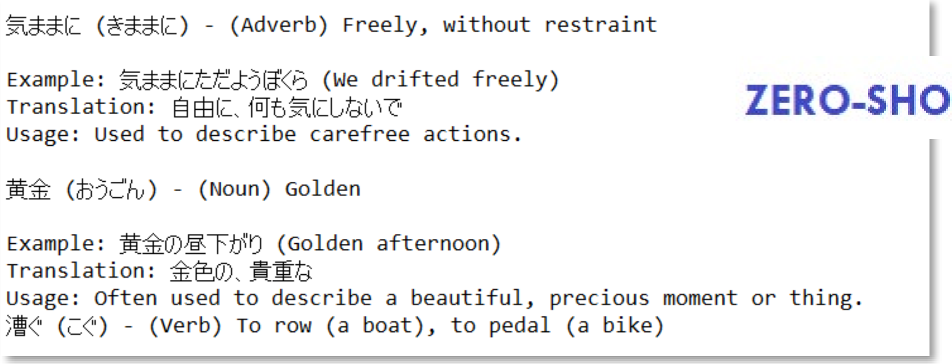
```
скучая – чувствам се отегчен, когато нямам какво да правя.
Пример: Алиса се умори да седи без работа.

скамейка – пейка, на която се седи, обикновено на открито или в парк.
Пример: Алиса седеше до сестра си на скамейката.

картинки – илюстрации или рисунки в книга.
Пример: Книгата, която четеше сестра й, нямаше картинки.

разговори – диалози или комуникация между хора.
Пример: Книгата не съдържаше нито картинки, нито разговори.

Бял Заек – герой от историята, необикновен заек с човешки качества.
Пример: Алиса видя Белия Заек да подскача край нея.
```

One-Shot

√ smaller number of words

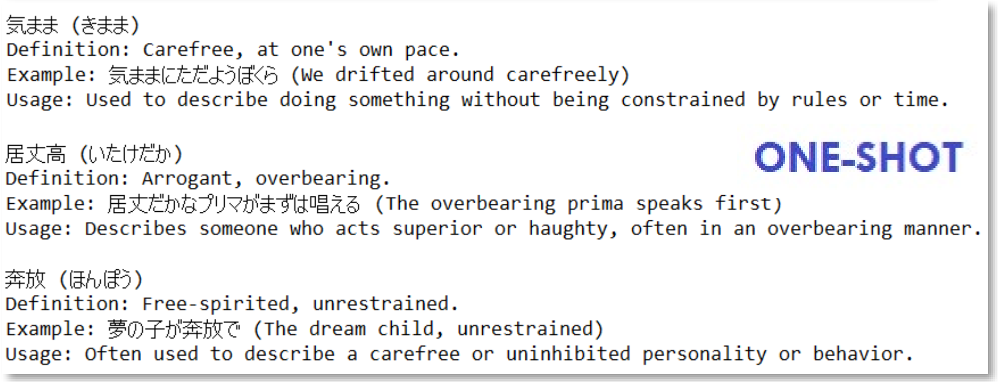
√ higher average frequency rank

Which Scenario Provided Better Results?



ZERO-SHOT

```
気ままに (きままに) - (Adverb) Freely, without restraint

Example: 気ままにただようぼくら (We drifted freely)
Translation: 自由に、何も気にしないで
Usage: Used to describe carefree actions.

黄金 (おうごん) - (Noun) Golden

Example: 黄金の昼下がり (Golden afternoon)
Translation: 金色の、貴重な
Usage: Often used to describe a beautiful, precious moment or thing.
漕ぐ (こぐ) - (Verb) To row (a boat), to pedal (a bike)
```



ONE-SHOT

```
気まま (きまま)
Definition: Carefree, at one's own pace.
Example: 気ままにただようぼくら (We drifted around carefreely)
Usage: Used to describe doing something without being constrained by rules or time.

居丈高 (いたけだか)
Definition: Arrogant, overbearing.
Example: 居丈だかなプリマがまずは唱える (The overbearing prima speaks first)
Usage: Describes someone who acts superior or haughty, often in an overbearing manner.

奔放 (ほんぽう)
Definition: Free-spirited, unrestrained.
Example: 夢の子が奔放で (The dream child, unrestrained)
Usage: Often used to describe a carefree or uninhibited personality or behavior.
```

One-Shot

√ no Japanese translations of the definitions

√ higher average frequency rank

Chain-of-Thought Experiment

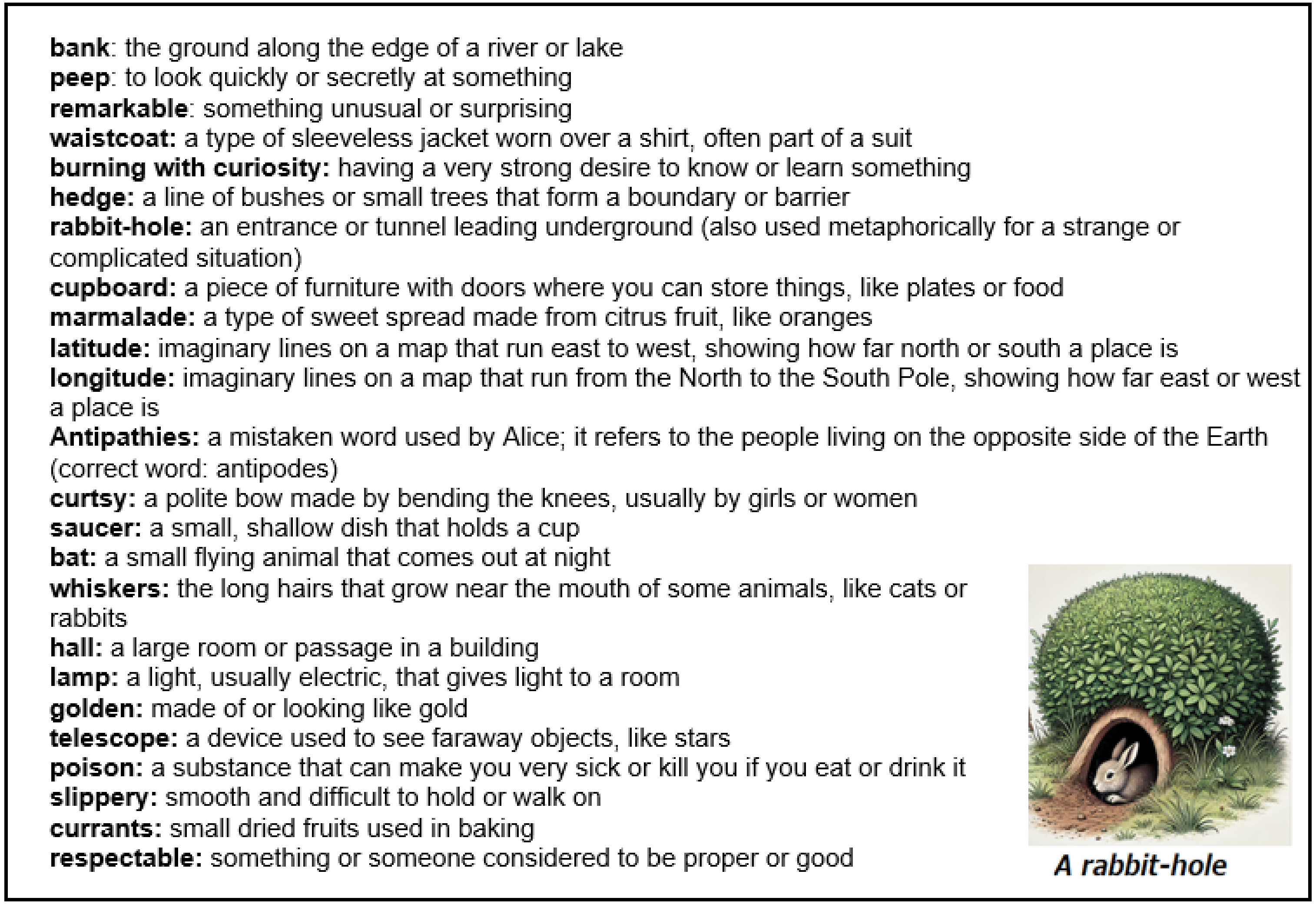
- performed on the one-shot English output
- prompt:

Please make the following changes to the glossary:

1. Remove the POS. Please keep them only in the case of items for which there is associated polysemy (e.g. ‘empty’ can be an adjective or a verb)
2. Add 1-2 expressions (i.e. items that are longer than a single word) to the glossary if you find suitable ones in the text. Exchanging a word from the glossary with an expression or combining words into an expression is also possible.
3. Provide an image to illustrate a suitable item from the glossary.

Please do not change anything else.

Chain-of-Thought Experiment: output



bank: the ground along the edge of a river or lake
peep: to look quickly or secretly at something
remarkable: something unusual or surprising
waistcoat: a type of sleeveless jacket worn over a shirt, often part of a suit
burning with curiosity: having a very strong desire to know or learn something
hedge: a line of bushes or small trees that form a boundary or barrier
rabbit-hole: an entrance or tunnel leading underground (also used metaphorically for a strange or complicated situation)
cupboard: a piece of furniture with doors where you can store things, like plates or food
marmalade: a type of sweet spread made from citrus fruit, like oranges
latitude: imaginary lines on a map that run east to west, showing how far north or south a place is
longitude: imaginary lines on a map that run from the North to the South Pole, showing how far east or west a place is
Antipathies: a mistaken word used by Alice; it refers to the people living on the opposite side of the Earth (correct word: antipodes)
curtsy: a polite bow made by bending the knees, usually by girls or women
saucer: a small, shallow dish that holds a cup
bat: a small flying animal that comes out at night
whiskers: the long hairs that grow near the mouth of some animals, like cats or rabbits
hall: a large room or passage in a building
lamp: a light, usually electric, that gives light to a room
golden: made of or looking like gold
telescope: a device used to see faraway objects, like stars
poison: a substance that can make you very sick or kill you if you eat or drink it
slippery: smooth and difficult to hold or walk on
currants: small dried fruits used in baking
respectable: something or someone considered to be proper or good

A rabbit-hole

Chain-of-Thought Experiment: evaluation

- an expression is added (‘burning with curiosity’)
- a relevant high-quality image is generated
- parts of speech are removed

👍

- POS are removed even in cases of polysemy (‘bank’)

👎

Conclusions and Future Work

- with professional glosses in mind, ChatGPT performs better at generating glossaries based on a learner-friendly text in a one-shot scenario
- further prompting significantly improves the output
- generation of glossaries by LLMs is promising and lower-resource languages can benefit from it
- the corpus needs to be increased to allow for better generalisation
- possible future combination with automatic text adaptation

Thank you for your kind attention!

Questions are welcome.

Paper:

SCAN ME

iglika.nikolova.stoupak@gmail.com

http://stih-sorbonne-universite.fr/

linkedin.com/in/iglika-nikolova-stoupak-71290029/