

ПРОЕКТ №1

АНАЛИЗ ОНОМАТОПОЭТИЧЕСКОЙ ЛЕКСИКИ В ПРОИЗВЕДЕНИЯХ А. П. ЧЕХОВА НА ЯПОНСКОМ ЯЗЫКЕ

Работу выполнила
Рыбакова Екатерина
2025 г.

Материал: 27 рассказов, опубликованных на портале «Аодзора-бунко», в переводе Киёси Дзиндзая (24 рассказа), Миэкити Судзуки (2 рассказа) и Каё Сэнумы (2 рассказа).

Ономатопеэтические единицы японского языка — это слова, непосредственно передающие звуки живой и неживой природы, физические и эмоциональные ощущения, описывающие действия и состояния предметов.



Этапы работы:



Часть 1. Создание списка ономатопоэтической лексики.

Часть 2. Парсинг произведений А. П. Чехова на японском языке с использованием BeautifulSoup и их сохранение в текстовых файлах.

Часть 3. Применение токенизатора японского языка `janome` со встроенным словарём и частеречной разметкой. Поиск ономатопоэтических единиц в текстах. Визуализация результатов (гистограммы, таблицы, облако слов и круговая диаграмма).

Часть 4. Более точная классификация ономатопов по частям речи при помощи грамматических правил. Визуализация результатов (таблицы, облака слов, круговая диаграмма). Выводы.

Часть 1. Создание списка ономатопоэтической лексики

1.1. Japanese_Manga_SFX.xlsx → .csv (576 слов).

1.2. giongo.txt → .csv (6148 слов). Сохранение общего файла со 3598 ономатопоэтическими словами (после удаления дублей) — combined_no_duplicates.csv.

1.3. Загрузка и подготовка csv-файла к работе (список onomatoroeia_list).

625	くさくさ	1593	ちゃかちゃか	3515	バラ
626	クサクサ	1594	チャカチャカ	3516	ばら
627	くしゃ	1595	ちやくちやく	3517	カッキー
628	クシャ	1596	チャクチャク	3518	かっきーん
629	ぐしゃ	1597	ちゃっかり	3519	カコーン
630	グシャ	1598	チャツカリ	3520	かこーん
631	くしゃくしゃ	1599	ちゃっぼちゃっぼ	3521	カクカク
632	クシャクシャ	1600	チャツポチャツポ	3522	かくかく
633	ぐしゃぐしゃ	1601	ちゃぶ	3523	びりっと
634	グシャグシャ	1602	チャブ	3524	びるりらびるりら
635	ぐじゃぐじゃ	1603	ちゃぼちゃぼ	3525	カボリ
636	グジャグジャ	1604	チャボチャボ	3526	かぼり
637	ぐしゃり	1605	ちやほや	3527	ばしっ
638	グシャリ	1606	チャホヤ	3528	パシュ

Применённые навыки:

- 1) Работа с pandas и таблицами .xlsx;
- 2) Регулярные выражения для поиска японских символов (re.compile);
- 3) Сохранение файла через функцию open()

Часть 2. Парсинг с использованием BeautifulSoup

作家名：

作家名読み：

ローマ字表記：

生年：

没年：

人物について：

チェーホフ アントン

チェーホフ アントン

Chekhov, Anton

1860-01-17

1904-07-02

 [「アントン・チェーホフ」](#)

公開中の作品

1. [イオヌイチ](#)

(新字新仮名、作品ID：43647)

→[神西 浩](#)(翻訳者)

2. [犬を連れ戻した奥さん](#)

(新字新仮名、作品ID：43644)

→[神西 浩](#)(翻訳者)

3. [かき](#)

(新字新仮名、作品ID：51292)

→[神西 浩](#)(翻訳者)

4. [カシタンカ](#)

(新字新仮名、作品ID：51366)

→[神西 浩](#)(翻訳者)

5. [かちめ](#)

——喜劇 四幕——(新字新仮名、作品ID：51860)

→[神西 浩](#)(翻訳者)

6. [可愛い女](#)

(新字新仮名、作品ID：43645)

→[神西 浩](#)(翻訳者)

7. [グーセフ](#)

(新字新仮名、作品ID：51424)

→[神西 浩](#)(翻訳者)

8. [霧の上のアナト](#)

(新字新仮名、作品ID：52177)

→[神西 浩](#)(翻訳者)

9. [麗](#)

笑劇 一幕(新字新仮名、作品ID：52239)

→[神西 浩](#)(翻訳者)

10. [法蘭](#)

(新字新仮名、作品ID：45763)

→[神西 浩](#)(翻訳者)

11. [子守つ子](#)

(新字旧仮名、作品ID：45196)

→[鈴木 三重吉](#)(翻訳者)

12. [桜の園](#)

(新字新仮名、作品ID：43598)

→[神西 浩](#)(翻訳者)

13. [少年たち](#)

(新字新仮名、作品ID：51352)

→[神西 浩](#)(翻訳者)

14. [小波瀾](#)

(新字新仮名、作品ID：51353)

→[神西 浩](#)(翻訳者)

15. [接吻](#)

(新字新仮名、作品ID：45762)

→[神西 浩](#)(翻訳者)

16. [大ヴォローヂャと小ヴォローヂャ](#)

(新字新仮名、作品ID：51389)

→[神西 浩](#)(翻訳者)

17. [道放されて](#)

(新字新仮名、作品ID：51735)

→[神西 浩](#)(翻訳者)

18. [妻](#)

(新字新仮名、作品ID：51736)

→[神西 浩](#)(翻訳者)

19. [てがみ](#)

(新字旧仮名、作品ID：45195)

→[鈴木 三重吉](#)(翻訳者)

20. [天才](#)

(新字新仮名、作品ID：52293)

→[神西 浩](#)(翻訳者)

21. [富蔵](#)

(新字新仮名、作品ID：51354)

→[神西 浩](#)(翻訳者)

22. [女房ども](#)

(新字新仮名、作品ID：51423)

→[神西 浩](#)(翻訳者)

23. [ぬい](#)

(新字新仮名、作品ID：51365)

→[神西 浩](#)(翻訳者)

24. [マリ・デル](#)

(新字新仮名、作品ID：52294)

→[神西 浩](#)(翻訳者)

25. [逢入り支度](#)

(新字新仮名、作品ID：51293)

→[神西 浩](#)(翻訳者)

26. [六号室](#)

(旧字旧仮名、作品ID：45667)

→[瀬沼 夏重](#)(翻訳者)

27. [六号室](#)

(新字新仮名、作品ID：50220)



→[瀬沼 夏重](#)(翻訳者)

28. [フーニャ伯父さん](#)

——田園生活の情景 四幕——(新字新仮名、作品ID：51862)

→[神西 浩](#)(翻訳者)

ファイルのダウンロード

ファイル種別	圧縮	ファイル名 (リンク)	文字集合/符号化方式	サイズ	初登録日	最終更新日
 テキストファイル(ルビあり)	zip	45195_nuby_19074.zip	JIS X 0208/ShiftJIS	4175	2005-08-19	2005-08-19
 XHTMLファイル	なし	45195_19224.html	JIS X 0208/ShiftJIS	9771	2005-08-19	2005-08-19

てがみ

アントン・チエーホフ Anton Chehov

鈴木三重吉訳

ユウコフは年はまだやつと九つです。せんには、お母さんと一しよに、あなかの村のマカリツチさまといふ、だんなのうちにおいてもらつてゐました。お母さんはそのうちの女中になつて、はたらいてゐたのです。そのお母さんが死んでしまつたので、ユウコフもそのお家にあらなくなり、人の世話で、三月まへから、この酒屋の店へ、奉公にはいつたのでした。

こよひはクリスマスの晩です。ユウコフは親方や同僚子たちが、教金からかへつてくるまでは、どんなにおくまでも、ねむらないで、まつてゐなければならぬのです。ユウコフは一人ぼつちで、さびしくてたまらないので、戸たなから、そつと親方のインキつぽをもちだして、さきのさゝくれたペンで、しわくちやの紙へてがみをかきました。だれよりも大すきな、あのマカリツチのだんなへ出さうとおもひついたので、

「コンスタンチン、マカリツチのだんなさま、」とユウコフは、くひをひねり／＼かき出しました。

「だんなさまのところは、クリスマスでにぎやかでせう。神さまが、だんなさまに、どうきりい／＼ことを下さるやうにいのつてます。だつて、わたしには、お父つあんもお母さんもみなくなつたから、あとは、たゞだんなさまだけです。」

ユウコフはこゝまでかくと、目をあげて、くらしい窓を見上げました。ろうそくの、くらしいあかりが、ガラスにほんやりうつてゐます。それをじつとみてみると、マカリツチのだんなの妻が、ありありとガラスの中にかひ上つて来ました。マカリツチのだんなは、年は六十五です。せいのひくい、やせた、それでゐてとても元気なおぢいさまです。いつもたのしさににこ／＼してばかりゐます。昼間は給所になごらんで、料理人をからかつたりしてゐますが、夜になると、大きな羊の毛皮の外とうにくるまつて、家裏の見まはりに出ていきます。だんなのうしろには、いつも、カシュタンカとエールといふ、二匹きの犬がついてゐます。

今だんなはどうしてゐるかしらとユウコフは思ひました。村の教会の窓はあか／＼とがすやいてゐるでせう。だんなはうちの門のところに、フェルトの靴をばた／＼させながら立つてゐて、女中たちをわらはせてゐるかもしれません。

「どうだい、一つかゝねえか、」

だんなは、かき煙草の箱を女中にわたします。女中はうけとつてかいてみて、とてもうれしがつて、はッはッとおもふのでした。

「くさいだらう。あとで鼻の先をよくふくんだぞ。——おゝお、ひどく、いてつくやねえか。おもしろくつくだらう。」

それから、だんなは、かき煙草を犬にもかゝせます。カシュタンカは、鼻をクン／＼ならして、にげだします。エールはむやみに尻尾をふつて、かすせないでくれといふやうにおせじをつかひます。

夜の空は、ふかく水色にはれて、村金太いがかつくりとうがかひ上つてゐます。まつ白に雪をかぶつた屋根や、煙をはいてる煙突やしもで銀色になつた木立などが、幻燈のやうにすんでゐます。空には、お星さまがおどけたやうにまたゝいてゐます。星の大河も、クリスマスがきたので、雪でみがきをかけたやうに、白くはつきりと光つてゐます。

ユウコフはためいきをして、またかきつめました。

Итого: 28 текстовых файлов.
У «Палаты №6» два издания: старое и новое.

Часть 3. Токенизатор janome

Janome — Python-библиотека для обработки японского текста, основанная на инструменте для морфологического анализа MeCab.

Благодаря встроенной POS-разметке можно легко избавиться от знаков препинания (記号) и стоп-слов:

- частицы (助詞): の, と, が, を, は и т.д.
- вспомогательные глаголы (助動詞): ~ます, ~である и т.д.
- союзы (接続詞): そして, それで и т.д.
- непредикативные, или приимённые, прилагательные (連体詞): この, その и т.д.

```
from janome.tokenizer import Tokenizer

tokenizer = Tokenizer()
text = "今のあなた方の話はすっかり聞いしまったの"
tokens = tokenizer.tokenize(text)

for token in tokens:
    print(f"Токен: {token.surface}")
    print(f"Словарная форма: {token.base_form}")
    print(f"Часть речи: {token.part_of_speech}")
    print(f"Чтение токена на катакане: {token.reading}")
    # Время, наклонение, число и др. (для прилагательных и глаголов)
    print(f"Форма слова: {token.infl_type}, {token.infl_form}")
    print("-" * 30)
```

✓ 0.3s

Токен: すっかり
Словарная форма: すっかり
Часть речи: 副詞, 一般, *, *
Чтение токена на катакане: スッカリ
Форма слова: *, *

Токен: 聞い
Словарная форма: 聞く
Часть речи: 動詞, 自立, *, *
Чтение токена на катакане: キイ
Форма слова: 五段・カ行イ音便, 連用タ接続

Токен: ちまつ
Словарная форма: ちまう
Часть речи: 動詞, 非自立, *, *
Чтение токена на катакане: チマツ
Форма слова: 五段・ワ行促音便, 連用タ接続

Токен: た
Словарная форма: た
Часть речи: 助動詞, *, *, *
Чтение токена на катакане: タ
Форма слова: 特殊・タ, 基本形

Однако поскольку ономапоэтические слова выступают в роли наречий (ADV, 副詞), глаголов (VERB, 動詞) и прилагательных (ADJ, 形容詞), мы сосредоточимся только на тех частях речи, которые необходимы для нашего анализа.

Также добавим к списку существительные (名詞), т.к. N-ADJ выполняют функцию прилагательного, но с грамматикой, свойственной именам существительным.

Дополнительно исключим из обработки зависимые слова (非自立) в additional_info существительных (в результатах на правах существительного присутствовала падежная частица の).

Usages	
ADV	• Describing an action or process (Adverb function) おんな こ わら 女の子はにこにこと笑った。The girl smiled.
する	• Used with する as a verb unit (と omitted) しんぞう 心臓がどきどきしている。My heart is pounding.
ADJ	• Describing a state of being (Adjective function) ごつごつ [とした] て している 手 : a rugged, sturdy hand
N-ADJ	• Adjective function, but with noun-like grammar. だぶだぶのズボン : baggy pants びしょびしょになる : to become soaking wet

Источник: nihonshock.com

Подключаем `onomatopoeia_list...`

...и ищем перечисленные единицы.

Выбор слов длиннее двух слогов обусловлен желанием повысить точность анализа и уменьшить количество ложноположительных результатов (в частности, из-за омонимии глагол `する` появляется больше 5000 раз). Длинные слова (**три слога и более**) с большей вероятностью представляют собой ономатопэтические единицы.

Кроме того, у нас появился своеобразный список стоп-слов, искажающих результаты, который тоже требует исключения из-за омонимии:

```
stoplist = ["ちゃん", "ありゃ", "そりゃ", "キュー", "ぐるり", "どっか"]
```

Итак, всего в 28 текстах было обнаружено **1495 ономатопэтических единиц**.

Почему бы не подсчитать долю ономатопэтических слов в каждом рассказе и в целом?

В рассказе `かき.txt` найдено 29 ономатопэтических слов:

はっきり (副詞), がくがく (副詞), ぐったり (副詞), だぶだぶ (名詞), ぼろぼろ (名詞), ちょう

В рассказе `かもめ.txt` найдено 119 ономатопэтических слов:

てくてく (副詞), ごろごろ (副詞), ザーッ (副詞), てっきり (副詞), べったり (副詞), きっかり

В рассказе `てがみ.txt` найдено 3 ономатопэтических слов:

ぼんやり (副詞), ありあり (副詞), みしみし (副詞)

В рассказе `ねむい.txt` найдено 38 ономатопэтических слов:

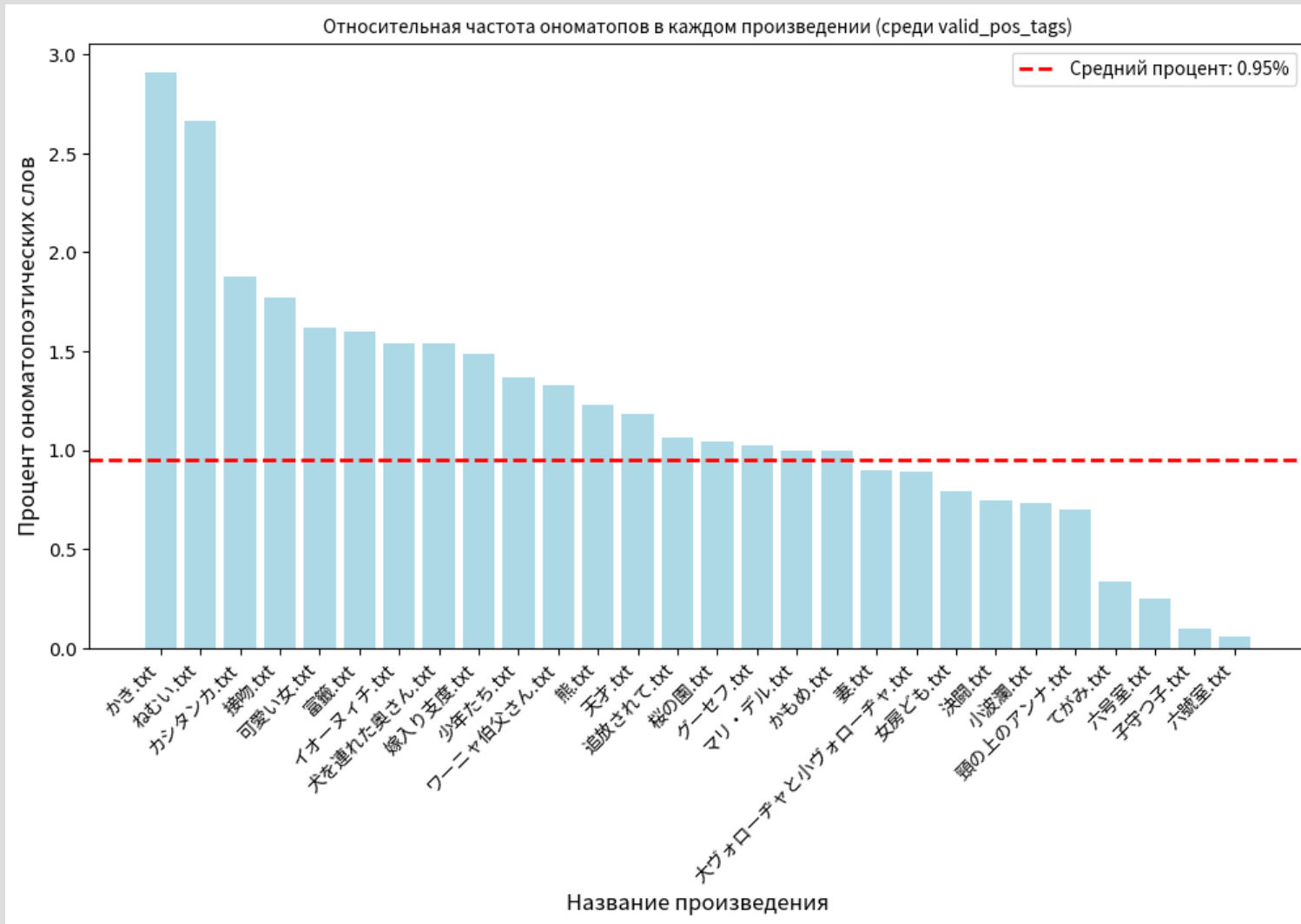
ころり (名詞), むんむん (副詞), すきずき (副詞), かさかさ (名詞), ころり (名詞), ぶつぶつ

В рассказе `イオーヌイチ.txt` найдено 79 ономатопэтических слов:

ちよいちよい (副詞), でっぷり (副詞), どっさり (副詞), さっぱり (副詞), ぶんぶん (副詞), ほ

И так далее...

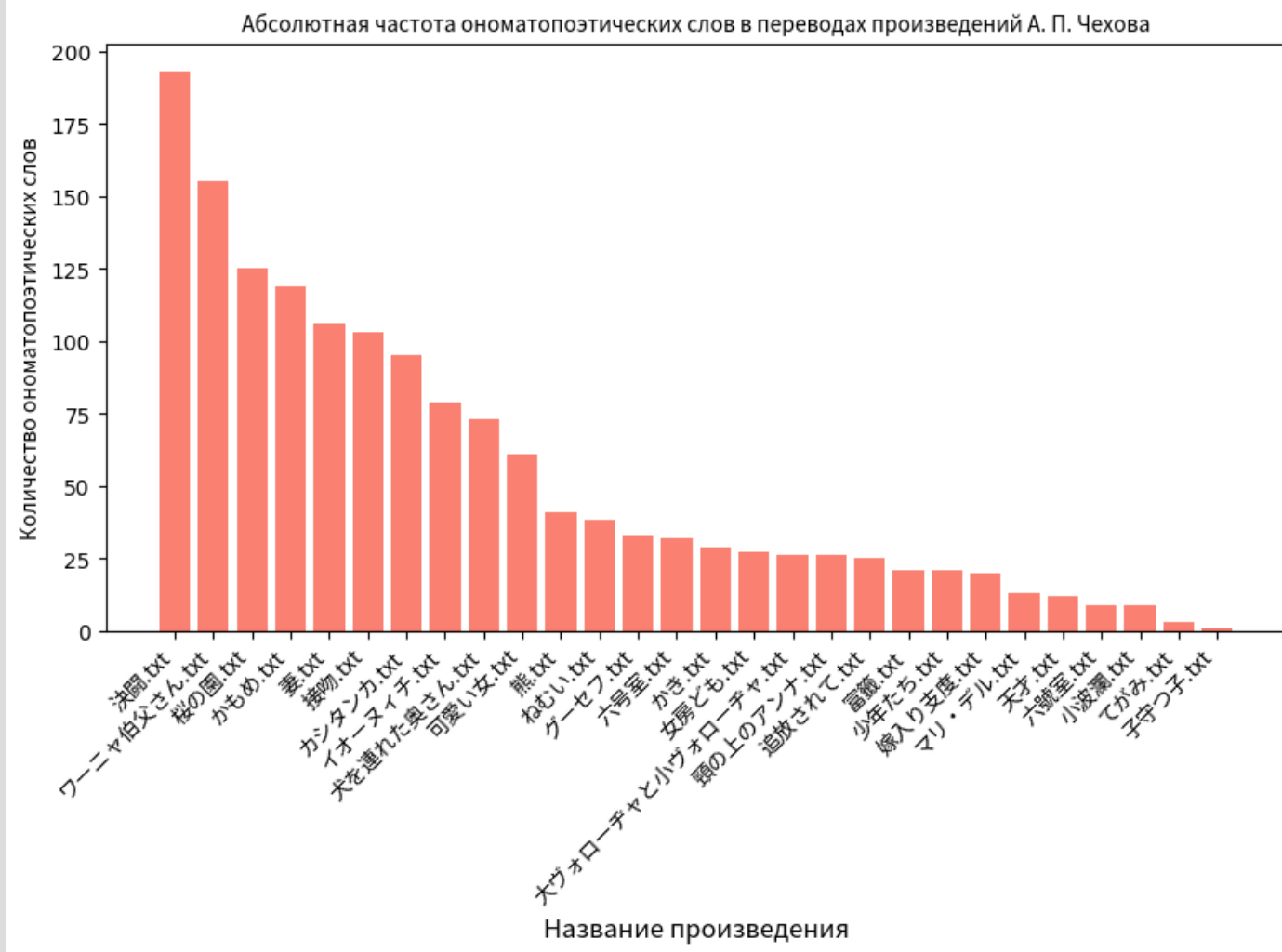
Всего найдено 1495 ономатопэтических слов.



Средний процент: 0,95%

Наибольший процент среди valid_pos_tags ономатопозитическая лексика занимает в рассказе «Устрицы».

Наименьший процент ономатопозитическая лексика занимает в старой версии повести «Палата №6».



Рейтинг по убыванию:

- 1) Повесть «Дуэль»: 194
- 2) Пьеса «Дядя Ваня»: 157
- 3) Пьеса «Вишнёвый сад»: 126
- 4) Пьеса «Чайка»: 119
- 5) Рассказ «Жена»: 106
- 6) Рассказ «Поцелуй»: 95
- 7) Рассказ «Каштанка»: 95
- 8) Рассказ «Ионыч»: 79
- 9) Рассказ «Дама с собачкой»: 73
- 10) Рассказ «Душечка»: 61
- 11) Пьеса «Медведь»: 41
- 12) Рассказ «Спать хочется»: 38
- 13) Рассказ «Гусев»: 33
- 14) Повесть «Палата №6»: 32 (нов. вер.)
- 15) Рассказ «Устрицы»: 29
- 16) Рассказ «Бабы»: 27
- 17) Рассказ «Володя большой и Володя маленький»: 26
- 18) Рассказ «Анна на шее»: 26
- 19) Рассказ «В ссылке»: 25
- 20) Рассказ «Выигрышный билет»: 21
- 21) Рассказ «Мальчики»: 21
- 22) Рассказ «Приданое»: 20
- 23) Рассказ «Mari d'elle»: 13
- 24) Рассказ «Талант»: 12
- 25) Повесть «Палата №6» (стар. вер.): 9
- 26) Рассказ «Житейская мелочь»: 9
- 27) Рассказ «Письмо»: 3
- 28) Рассказ «Ванька»: 1

[illegible]

Единица	Частотность	Часть речи	Перевод
0 すっかり	121	副詞	Полностью
1 はっきり	46	副詞	четко
2 ちょうど	46	副詞	только
3 さっぱり	40	副詞	Освежающий
4 そろそろ	33	副詞	Пришло время
5 うとうと	21	副詞	Засыпаю
6 どっさり	18	副詞	Много
7 つくづく	18	副詞	Действительно
8 ぶらぶら	18	副詞	Блуждание вокруг
9 わざわざ	18	副詞	Специально
10 ちらちら	17	副詞	мерцание
11 そっくり	17	副詞	Точно
12 へとへと	17	名詞	усталый
13 きらきら	16	副詞	Игристое
14 うっとり	16	副詞	Зачарованный
15 しっかり	15	副詞	Твердо
16 ゆっくり	15	副詞	медленно
17 おずおず	14	副詞	Робко
18 ふらふら	14	副詞	Неустойчивый
19 にこにこ	13	副詞	Улыбаясь

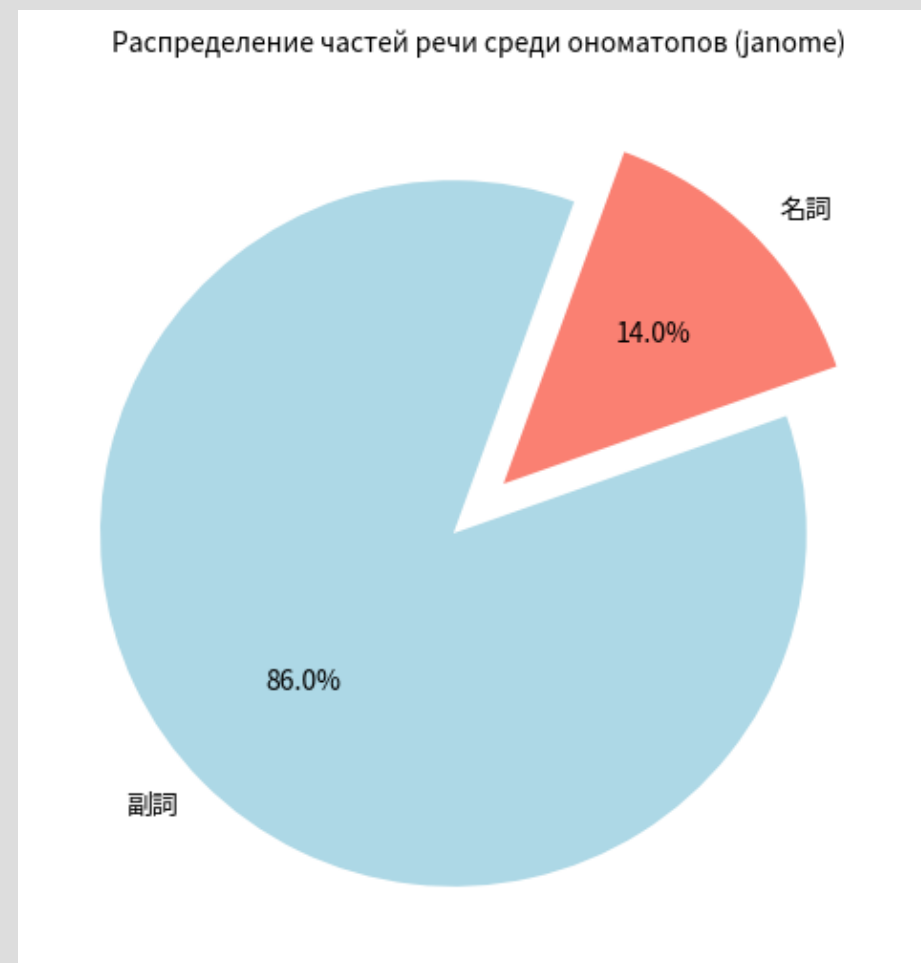
```
# from deep_translator import GoogleTranslator
(по независящим причинам может не работать)
```

Распределение частей речи среди ономатопов (janome)

Наречия (86%), существительные (14%).

Результаты поиска отвечают поставленным задачам рассмотреть частотность ономатопов, однако ограниченный функционал частеречной разметки janome не позволяет выявить глаголы и прилагательные, а также порой ошибочно относит наречия к существительным.

В следующей части проекта мы постараемся более точно классифицировать части речи ономатопоэтических слов при помощи правил японской грамматики.



Часть 4. Грамматические правила для определения части речи

- ❖ **Глаголы и глагольные конструкции (adjective function):** частица *ga* (опционально) + ядро + глагол с начальной формой *суру*.
- ❖ **Прилагательные (N-ADJ):** ядро + частица *но*, частица *ни*, суффикс *на* или глагольные частицы *да* | *дэсу*.
- ❖ **Наречия:** ядро + частицы *то* | *дэ*. С большой вероятностью всё остальное, что не подходит под условия выше.

Нами было принято решение присоединить ономастои формата **ядро + глагол *сита* (した) или *ситэиру* (している)** к глаголам, а не к прилагательным, исходя из особенностей POS-разметки, предоставляемой *janome*. В данном случае *janome* разъединяет глагол *сита* на составляющие *си* и *та*, что осложняет поиск слов в контексте.

Поэтому, для упрощения обработки текстов в текущем проекте, было принято решение считать такие конструкции глагольными, так как они фактически представляют собой комбинации глагола и прилагательного. Такой шаг позволяет более эффективно обрабатывать подобные формы, не слишком усложняя процесс разметки и поиска.

Результаты для файла *かき.txt*:

VERB: *がくがく*, *ちらちら*, *へどもど*, *つるつる*,
ADV: *はつきり*, *ぐったり*, *ちょうど*, *てくてく*,
N-ADJ: *だぶだぶ*, *ぼろぼろ*

Результаты для файла *かもめ.txt*:

VERB: *べたべた*, *さばさば*, *どきどき*, *やきもき*,
ADV: *てくてく*, *ごろごろ*, *ザーツ*, *てつきり*, *べ*
N-ADJ: *もじやもじや*, *そっくり*, *めちゃめちゃ*,

Результаты для файла *てがみ.txt*:

VERB: (не найдено)
ADV: *ぼんやり*, *ありあり*, *みしみし*
N-ADJ: (не найдено)

Результаты для файла *ねむい.txt*:

VERB: *むんむん*, *ずきずき*, *もやもや*, *むんむん*,
ADV: *ころり*, *ころり*, *ぶつぶつ*, *うとうと*, *よた*
N-ADJ: *かさかさ*, *へとへと*

Результаты для файла *イオーヌイチ.txt*:

VERB: *でっぶり*, *ぶんぶん*, *まるまる*, *うとうと*,
ADV: *ちょいちょい*, *どっさり*, *さっぱり*, *ほやほ*
N-ADJ: *とんちんかん*, *こちこち*, *こちこち*, *くた*

Распределение частей речи среди ономатопов (правила)

Наречия (70.4%), глаголы (23.4%),
прилагательные (6.3%).

Ономатопея в японском языке чаще всего используется для описания звуков, движений или состояний, что естественным образом делает её наречием — с чем мы и *janome* оказались согласны.



[illegible]

	Единица	Частотность	Перевод
0	すっかり	121	Полностью
1	ちょうど	46	только
2	さっぱり	37	Освежающий
3	はっきり	33	четко
4	そろそろ	32	Пришло время
5	やれやれ	22	о, Боже
6	つくづく	18	Действительно
7	わざわざ	18	Специально
8	どっさり	17	Много
9	ゆっくり	14	медленно
10	おずおず	12	Робко
11	ひそひそ	12	Шепот
12	ぶるぶる	11	дрожащий
13	ちょいちょい	11	изредка
14	ぐっすり	11	Спи спокойно
15	そっくり	10	Точно
16	がらん	10	Пустой
17	しっかり	10	Твердо
18	てっきり	9	Конечно
19	しみじみ	9	Глубоко

Топ-20 частотных онематопов среди глаголов



	Единица	Частотность	Перевод
0	うとうと	14	Засыпаю
1	はっきり	13	четко
2	きらきら	11	Игристое
3	ぶらぶら	11	Бродить вокруг
4	いらいら	11	Раздражительность
5	にこにこ	10	Улыбаясь
6	ぷりぷり	10	Упругий
7	ちらちら	9	мерцание
8	うっとり	9	Зачарованный
9	むんむん	9	душно
10	うんざり	8	сыт по горло
11	ぷんぷん	7	Злой
12	おどおど	6	Робко
13	じめじめ	6	влажный
14	もやもや	6	Размытость
15	わくわく	6	Захватывающий
16	そわそわ	6	беспокойный
17	にっこり	5	Улыбка
18	しっかり	5	Твердо
19	ぐずぐず	5	Дизеринг

Топ-20 частотных онематопов среди прилагательных



	Единица	Частотность	Перевод
0	へとへと	16	усталый
1	そっくり	7	Точно как
2	だぶだぶ	5	Мешковатый
3	くたくた	5	измученный
4	かさかさ	4	Сухость
5	たっぷり	4	Множество
6	ぼろぼろ	3	Потертый
7	ちよっぴり	3	Немного
8	ふらふら	3	Неустойчивый
9	よれよれ	3	Изношенный
10	ぎりぎり	3	Едва
11	めちゃめちゃ	2	Очень сильно
12	ぐらぐら	2	свободный
13	こちこち	2	Жесткий
14	ぶくぶく	2	Пузыристый
15	よぼよぼ	2	Йобойобо
16	カサカサ	2	Шуршание
17	もじゃもじゃ	1	Кустистый
18	ぱりぱり	1	хрустящий
19	さわやか	1	Освежающий

Выводы:

Таким образом, с помощью `janome` у нас получилось найти оноματοпоэтические слова из созданного заранее списка, однако встроенной POS-разметки оказалось недостаточно, чтобы определить часть речи оноματοпоэтической единицы в контексте.

Мы постарались вручную написать правила, позволяющие с большей точностью определять часть речи, и это значительно улучшило результаты. Теперь стало наглядно, какой процент занимают глаголы с глагольными выражениями, а какой — прилагательные.

Конечно, созданный нами алгоритм допускает ошибки. Несмотря на то что в рамках данного проекта нам не удалось отладить определение части речи до безупречности (в частности, мы столкнулись с задачей отсоединения глагольных выражений от глаголов), это может быть доработано в будущем.

Кроме того, было бы интересно изучить, как со временем изменялось употребление оноματοпоэтических слов в японском языке, а также рассмотреть труды японских исследователей о распределении частей речи среди оноματοпов.

Спасибо за внимание!