ABRÉGEMENT DES TEXTES LITTÉRAIRES DANS PLUSIEURS LANGUES : TRANSFORMATIONS GRAMMATICALES

IGLIKA NIKOLOVA-STOUPAK

Encadrement : Eva Lacroix, Gaël Lejeune
Année: 1

21/12/2023

MA THÈSE : PLAN PROVISOIRE

« Production de versions abrégées de textes littéraires : une approche multilingue »

I. Études bibliographiques
    1. Modèles de synthèse de l'IA
    2. Comparaison interlinguistique
II. Acquisition du langage
    1. Transformations grammaticales dans les versions abrégées de textes littéraires
    2. Abrégement de textes littéraires pour lecteurs autistes
III. Accent sur des langues spécifiques
    1. Les versions abrégées en bulgare sont-elles trop libres ?
    2. Le discours dans la série abrégée japonaise Aoi Tori Bunko
IV. Des machines et des hommes
    1. Évaluation des LLM de pointe dans le cadre de la tâche de synthèse de textes littéraires
    2. Abrégement automatique des textes dans des langues rares
V. Testes

J'AI COMMENCÉ PAR LA QUESTION...

Est-ce qu'il arrive d'AJOUTER du contenu lorsque nous composons des versions abrégées de textes ?

L'ACQUISITION D'ÉLÉMENTS DE VOCABULAIRE COMPLEXES

- Leung et al (2021) : Les parents introduisent des mots complexes en affinant leur langage et fournissant un contexte explicatif supplémentaire

(idées de « motherese language » et « foreigner talk » - Ferguson, 1977)

Le léopard pointillé courait après le lapin.

Le léopard courait après le lapin comme un chat.

- Christophe et al (2010): Les connaissances des différents éléments du langage (ex. lexique/phonologie/syntaxe) s'entraident mais aussi, paradoxalement, il faut connaître l'un pour comprendre l'autre.

  => un vocabulaire complexe doit-il donc être associé à une grammaire simple pour faciliter le processus d'acquisition ?

HYPOTHÈSE

Les versions abrégées de textes littéraires devraient introduire des éléments de vocabulaire complexes en simplifiant et en élargissant le contexte dans lequel ils apparaissent.

QU'EST-CE QU'UN MOT COMPLEXE ?

- sens : un mot abstrait ?
- morphologie : un mot qui a de nombreux affixes ?
- fréquence : un mot rarement utilisé ?
- longueur : un long mot ? (Loi du moindre effort de Zipf)

- Lors de mes premières expériences, j'ai utilisé la longueur.

QU'EST-CE QU'UN MOT COMPLEXE ? TRAVAUX FUTURS

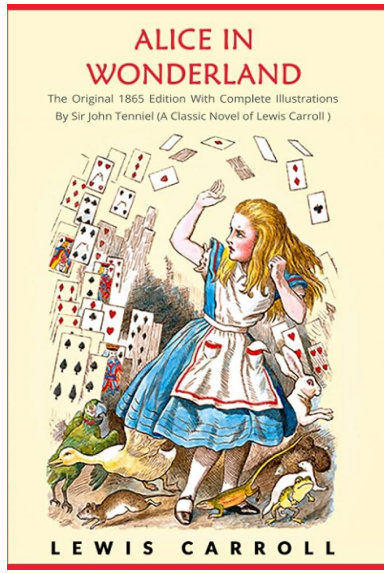
- consacrer une section à la question théorique de la complexité des mots
- utiliser différentes méthodes TAL pour dériver des listes de mots complexes (longueur, listes de fréquences, analyse morphologique) et voir quels mots apparaissent systématiquement comme complexes
- calculer le nombre de mots complexes par texte et par langue

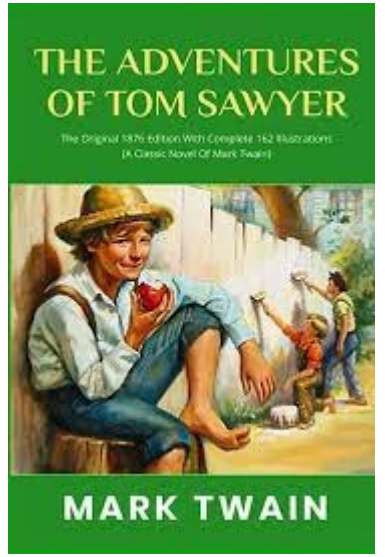
```
#English list
freq_list_en = pd.read_csv('/kaggle/input/pixel2023-frequency-lists/frequency_list_en.txt', sep='\t')
freq_list_en = freq_list_en.drop(['Rank', 'Count (per billion)'], axis=1)
freq_list_en = freq_list_en.Word.values.tolist() #turn dataframe into list
#remove apostrophies to make same format as list of repeated words in texts
for word in freq_list_en:
    word = word.replace("'","")

#French list
freq_list_fr = pd.read_csv('/kaggle/input/pixel2023-frequency-lists/frequency_list_fr.txt', names = ["Word"], sep='/n')
for i in range(0, len(freq_list_fr)):
    freq_list_fr.Word[i] = ''.join([i for i in freq_list_fr.Word[i] if not i.isdigit()])
    freq_list_fr.Word[i] = freq_list_fr.Word[i].replace(" . ","")
freq_list_fr = freq_list_fr.Word.values.tolist()

#Russian list (remove info in brackets after each word)
freq_list_ru = pd.read_csv('/kaggle/input/pixel2023-frequency-lists/frequency_list_ru.txt', names = ["Word"])
for i in range(len(freq_list_ru)):
    temp = freq_list_ru.Word[i].find("(")
    freq_list_ru.Word[i] = freq_list_ru.Word[i][:temp-1]
freq_list_ru = freq_list_ru.Word.values.tolist()

#Spanish list
freq_list_sp = pd.read_csv('/kaggle/input/pixel2023-frequency-lists/frequency_list_sp.txt', sep=' ')
with open('/kaggle/input/pixel2023-frequency-lists/frequency_list_sp.txt') as f:
    freq_list_sp = f.readlines()
freq_list_sp = " ".join(freq_list_sp)
freq_list_sp = freq_list_sp.split()
#remove caps apostrophies,spaces and hyphens to make same format as list of repeated words in texts
for word in freq_list_sp:
    word = word.replace("'","")
    word = word.replace(" ","")
    word = word.replace("-","")
    word = word.lower()
```

EXPÉRIENCES

CORPUS

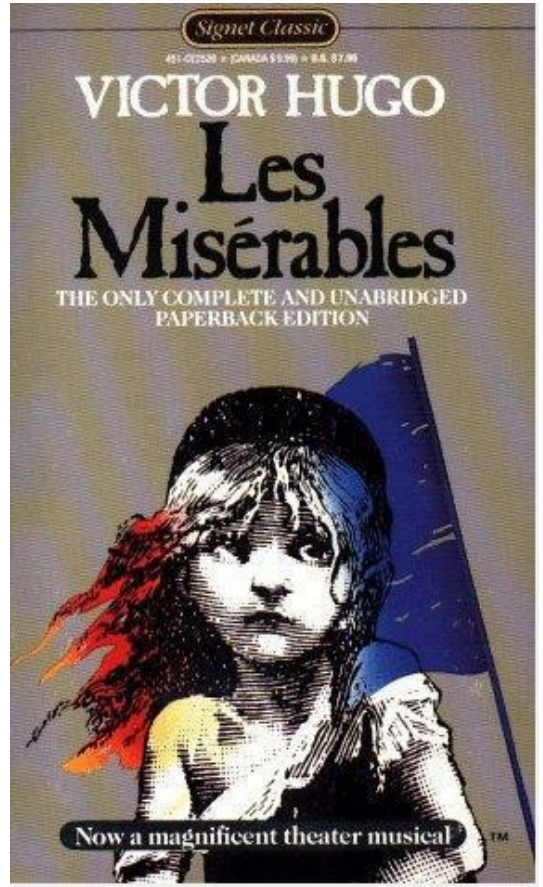
- Textes évalués : romans jeunesse célèbres (de la même époque)
  - Charles Dickens, *Un chant de Noël*
  - Lewis Carroll, *Alice au Pays des Merveilles*
  - Mark Twain, *Les aventures de Tom Sawyer*
- Langues
  - anglais
  - français
  - russe
  - espagnol
- Versions complètes et abrégées





CORPUS : TRAVAUX FUTURS

- Le corpus actuel est trop petit pour être généralisable (par exemple en termes de types de transformations ou de tendances par langue)
- Notamment, toutes les versions abrégées destinées aux apprenants de langues étrangères sont en anglais
- J'ai ajouté *Les misérables* en plusieurs versions, y compris pour les étudiants de français langue étrangère



PIPELINE DE TRAITEMENT AUTOMATIQUE

- Des fichiers .txt sont dérivés de tous les textes
- Les textes sont prétraités de manière standard
- Une liste des (300) noms les plus longs est établie pour chaque texte
- Pour chaque couple complet/abrégé, une liste est composée des noms complexes qui sont présents dans les deux textes
- Les contextes dans lesquels les mots apparaissent sont extraits et comparés
- La caractéristique recherchée est l'ajout de contexte dans la version abrégée

COMPLET : "and dashed the wild year through there stood a **solitary lighthouse** great heaps of seaweed clung to its base"

ABRÉGÉ : "crashed violently underneath them the spirit took scrooge to a **lighthouse built on a lonely rock** several miles from land"

ERREURS DÉTECTÉES

- non-mots issus d'un traitement textuel (« aventurescommencement »)
- mots marqués à tort comme noms (« aficionada », utilisé comme adjectif)
- mots qui ne sont pas complexes malgré leur longueur (« Christmas », « cumpleaños »)
- tandis que les collocations écrits avec tiret (« school-house ») sont volontairement incluses, les expressions en français et en russe qui contiennent des tirets en fonction de règles grammaticales (« commença-t-elle », « как-нибудь ») sont supprimées

RÉSULTATS

TYPES D'ADDITION DE CONTEXTE

- Au total, 62 instances de la caractéristique ont été trouvées dans 377 introductions parallèles de vocabulaire complexe.
- Ils ont été classés en 3 types :

1. ajout de vocabulaire connexe (46.8 %)

   "brouillard" vs "épais brouillard"

   "гостеприимство" vs "щедрое гостеприимство"

2. structure grammaticale plus simple (41.9 %)

   "tejer una guirnalda de margaritas" vs "juntar margaritas para trenzar una guirnalda"

3. explication ou définition (11.3 %)

   "в суде заседают, потому и называются 'присяжные заседатели'"

MOTS DU MÊME CHAMP LEXICAL

- Aident à comprendre le sens d'un mot et à l'utiliser ultérieurement
- Ils jouent un rôle clé dans les textes narratifs car ils se rapportent aux anaphores comme des synonymes (par exemple: « J'ai escaladé... » -- « le sommet... »)(Halliday, 2002)

MOTS DU MÊME CHAMP LEXICAL : TRAVAUX FUTURS

- Imiter la transformation en ajoutant des collocations aux mots complexes (garder à l'esprit que le sens peut être modifié)
- Utiliser l'outil *nltk.collocations* pour ajouter les collocations les plus courantes trouvées dans un grand corpus

TRANSFORMATIONS HARRISIENS ? HARRIS ET SES THÉORIES

- Zellig Harris est un linguiste du XXe siècle qui a travaillé sur la linguistique structurelle
- Il était professeur et ami de Chomsky
- Harris a cherché à systématiser le langage sans se référer à sa signification
- Le concept TAL de *plongement lexical* est basé sur sa théorie (Harris, 1988)

TRANSFORMATIONS HARRISIENS ?

- Les paires de phrases ont une « équivalence descriptive », c'est-à-dire qu'elles contiennent la même information
- Les versions plus courtes (ou « réductions ») provoquent une ambiguïté, tandis que la redondance facilite la compréhension
- Les types 1 et 3 de contexte ajouté ressemblent à la notion de « rapport » de Harris (description, définition)
- Le type 2 ressemble à des « paraphrases » (changements morphologiques et syntaxiques)

TRANSFORMATIONS HARRISIENS ? TRAVAUX FUTURS

- Harris donne des exemples plus précis de transformations grammaticales, telles que la conversion de la voix passive en voix active ou l'ajout d'une phrase du type « comme je l'ai dit »
- Ces simplifications ne sont pas nécessairement associées à un contexte plus long
- Imiter des types de transformations trouvés dans : la théorie de Harris (1988), Ibrahim (2002), les études du professeur Claire Martinot et de son équipe et dans le corpus
- Dériver automatiquement des versions abrégées des phrases

EXPLICATIONS EXPLICITES

- Jouent le même rôle que les glossaires ou les notes de bas de page qui fournissent des définitions (lorsqu'ils sont présents, ils peuvent également être ajoutés aux statistiques)

EXPLICATIONS EXPLICITES : TRAVAUX FUTURS

- Pour imiter la caractéristique, un dictionnaire peut être utilisé pour fournir des définitions de mots complexes (par exemple PyDictionary)

SIMPLIFICATIONS QUI N'IMPLIQUENT PAS DE CONTEXTE SUPPLÉMENTAIRE (PLUS DE TRAVAIL MANUEL REQUIS)

- Partie du discours atypique
  - exemple : « mangeable » vs « manger »
  - transformation (semi-automatique) : comparer la fréquence d'un mot avec la fréquence de son lemme ; s'il y a une grande différence, modifier la phrase
- Mots complexes supplémentaires :
  - exemple: « under sentence of execution » vs « in prison waiting for execution »
  - transformation : remplacez-les par un synonyme avec une plus haute fréquence en utilisant un thesaurus (noter qu'il ne s'agit plus de « transformations grammaticales »)
- Usage atypique d'un mot (idiosyncratique pour une langue)
  - exemple : *bien + ADJ ; NOM + important + PREP*
  - transformation : remplacer par *très + ADJ ; grand + NOM + PREP*

RÉSULTATS PAR LANGUE

Langue	Prévalence de la caractéristique (proportion)	Prévalence de la caractéristique (%)	Par type de contexte ajouté
anglais	4/69	12.2%	50% trans. gram.
			50% voc. ajouté
français	14/67	20.9%	28.6% trans. gram.
			42.9% voc. ajouté
			28.6% explication
russe	15/85	17.6%	46.7% trans. gram.
			33.3% voc. ajouté
			20% explication
espagnol	29/156	18.6%	44.8% trans. gram.
			55.2% voc. ajouté

RÉSULTATS PAR LANGUE

- Le nombre d'occurrences de la caractéristique est le plus faible dans les textes anglais et le plus élevé dans les textes français
- Les scores sont similaires pour le français et l'espagnol, ce qui est un résultat cohérent étant donné que les langues appartiennent à la même famille
- La répartition est équilibrée entre « transformation grammaticale » et « vocabulaire associé », un nombre limité d'explications apparaissant en français et en russe

RÉSULTATS PAR PUBLIQUE

Publique	Prévalence de la caractéristique (proportion)	Prévalence de la caractéristique (%)	Par type de contexte ajouté
générale/indéfinie	24/139	17.3%	50% trans. gram.
			41.7% voc. ajouté.
			8.3% explication
étudiants LE	4/15	26.7%	50% trans. gram.
			50% voc. ajouté
0-500 mots	*N/A*	*N/A*	*N/A*
500-1000 mots	*2/5*	*40%*	*50% trans. gram.*
			50% voc. ajouté.
enfants	28/169	16.6%	35.7% gram. trans.
			44.4% voc. ajouté
			17.9% explication
âge 5-8	*2/10*	*20%*	*100% explication*
âge 9-11	*14/95*	*14.7%*	*50% trans. gram.*
			50% voc. ajouté

RÉSULTATS PAR PUBLIQUE

- Il existe des tendances très similaires dans les textes destinés aux jeunes enfants et aux apprenants de niveau débutant
- Les mots complexes recherchés ont tendance à ne pas apparaître du tout dans les versions destinées aux très jeunes enfants ou aux apprenants de très faible niveau
- À l'inverse, la caractéristique est plus marquée chez les apprenants de langues étrangères de niveau supérieur (40 % des mots considérés).

LIMITATIONS

- Des œuvres pour le corpus sont parfois difficiles à trouver ou payantes
- Actuellement, les résultats ne sont pas assez généralisables (les choix des auteurs jouent un rôle trop important)
- La disponibilité des outils TAL pertinents est limitée (par exemple, dictionnaires dans des langues autres que l'anglais)
- L'abrégement automatique doit être limité au niveau de la phrase

TRAVAUX FUTURS : ANALYSE INTERLINGUISTIQUE

- La complexité du langage peut être volontaire ou inhérente à la langue (Ibrahim, 2002). Cette étude compare l'interaction entre la complexité de vocabulaire inhérente et la complexité grammaticale délibérément sélectionnée
- Calculer la distance cosinus entre les phrases originales et transformées par langue

IDÉE DE CONFÉRENCE

LxGr2024: Call for Papers

9th Symposium on Corpus Approaches to Lexicogrammar (LxGr2024)

5-6 July 2024 (online)

CALL FOR PAPERS

Deadline for abstract submission: Friday 15 March 2024

LxGr primarily welcomes papers reporting on **corpus-based research** on any aspect of the **interaction of lexis and grammar** — particularly studies that interrogate the system lexicogrammatically to get lexicogrammatical answers. However, **position papers** discussing **theoretical** or **methodological issues** are also welcome, as long as they are relevant to both lexicogrammar and corpus linguistics.

BIBLIOGRAPHIE

Christophe, A., Millotte, S., Brusini, P., Cauvet, E. (2010). Early Bootstrapping of Syntactic Acquisition. In Language Acquisition Across Linguistic and Cognitive Systems, 53-66.

DuBay, W. H. (2007), The Classic Readability Studies, Clearinghouse.

Gala, N., Todirascu, A., Bernhard, D., Wilkens, R. and Meyer, J.-P. (2020). Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés, Congrès Mondial de Linguistique Française, Montpellier, France.

Halliday, M. A. K. (2002). Linguistic Studies of Text and Discourse, Continuum.

Harris, Z. (1988). Language and Information, Columbia University Press.

Ibrahim, A. H. (2013). Une mesure unifiée de la complexité linguistique : l'analyse matricielle définitoire, Nouvelles perspectives en sciences sociales, 9: 17-80.

Leung, A., Tunkel, A. and Yurovsky, D. (2021). Parents Fine-Tune Their Speech to Children's Vocabulary Knoweldge, Psychological Science, 32: 975–984

Moje, E. B., Overby, M., Tysvaer, N. and Morris, K. (2008). The Complex World of Adolescent Literacy: Myths, Motivations, and Mysteries, Harvard Educational Review 78, 1: 107–54.

Rodriguez, A., Leonor, G. and Flórez, E. E. R. (2018). Using the Abridged Version of Some Novels as a Way to Encourage Students' Written and Oral Production, GiST Education and Learning Research Journal, 16: 6-32.

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K., and François, T. (2022) FABRA: French Aggregator-Based Readability Assessment toolkit, Proceedings of the Thirteenth Language Resources and Evaluation Conference, 1217–1233.

COPRUS 1/3

- Carroll, L. (1865). *Alice's Adventures in Wonderland*. Project Gutenberg.
- Carroll, L. (1916). *Alice's Adventures in Wonderland* (Abr. ed.). Sam'l Gabriel Sons and Company. (Original work published 1865)
- Carroll, L. (2000). *Alice's Adventures in Wonderland* (J. Bassett, Abr. ed.). Oxford University Press. (Original work published 1865)
- Carroll, L. (1978). *Alisa v Strane chudes*. Nauka (N. M. Demurova, Trans.). (Original work published 1865)
- Carroll, L. (1991). *Alisa v Strane chudes* (L. Yahnin, Abr. ed.). Eksmo. (Original work published 1865)
- Carroll, L. (2000). *Alisa v Strane chudes*. Biblioteka Maksima Moshkova (Y. Nesterenko, Trans.). (Original work published 1865)
- Carroll, L. (2018). *Alisa v Strane chudes* (Abr. ed.). Eksmo. (Original work published 1865)
- Carroll, L. (1996). *Las aventuras de Alicia en el país de las maravillas* (L. Maristany, Trans.). Titivillus. (Original work published 1865)
- Carroll, L. (2003). *Las aventuras de Alicia en el país de las maravillas* (M. Aguirre, Trans.). Ediciones del Sur. (Original work published 1865)
- Carroll, L. (2017). *Las aventuras de Alicia en el país de las maravillas* (N. Schuff, Abr. ed.). Santa Fe. (Original work published 1865)
- Carroll, L. (2018). *Las aventuras de Alicia en el país de las maravillas* (F. Díez de Miranda, Abr. ed.). Zig Zag. (Original work published 1865)
- Carroll, L. (1908). *Les Aventures d'Alice au pays des merveilles* (H. Bué, Trans.). Hachette. (Original work published 1865)
- Carroll, L. (1975). *Les Aventures d'Alice au pays des merveilles* (H. Parisot, Abr. ed.). Editions Corentin. (Original work published 1865)
- Carroll, L. (1992). *Les Aventures d'Alice au pays des merveilles* (P. Rouard, Abr. ed.). Bayard Jeunesse. (Original work published 1865)
- Carroll, L. (2001). *Les Aventures d'Alice au pays des merveilles* (J. Papy, Trans.). Gallimard Jeunesse. (Original work published 1865)
- Carroll, L. (2012). *Les Aventures d'Alice au pays des merveilles* (P. Protet, Abr. ed.). Auzou. (Original work published 1865)

COPRUS 2/3

- Dickens, C. (1905). *A Christmas Carol*. The Baker & Taylor Company.
- Dickens, C. (2000). *A Christmas Carol* (C. West, Abr. ed.). Oxford University Press. (Original work published 1905)
- Dickens, C. (2004). *A Christmas Carol* (P. Lagendijk, Abr. ed.). Mediasat Poland Bis. (Original work published 1905)
- Dickens, C. (1986). *Canción de Navidad* (S. R. Santerbás, Abr. ed.). Anaya. (Original work published 1905)
- Dickens, C. (2004). *Canción de Navidad* (Trans.). Ediciones del Sur. (Original work published 1905)
- Dickens, C. *Canción de Navidad* (Abr. ed.). Ediciones la Cueva, https://www.argentina.gob.ar/sites/default/files/dickens_charles_-_una_cancion_de_navidad.pdf. (Original work published 1905)
- Dickens, C. (1890). *Conte de Noël* (A. De Goy and De Saint-Romain, Trans.). La Bibliothèque électronique du Québec. (Original work published 1905)
- Dickens, C. (2002). *Conte de Noël* (Trans.). Pitbook.com, https://www.pitbook.com/textes/htm/chant_noel.htm. (Original work published 1905)
- Dickens, C. (2021). *Cuento de Navidad* (Abr. ed.). Blurb, https://www.studocu.com/es-mx/document/universidad-autonoma-agraria-antonio-narro/agricultura-sustentable/charles-dickens-cuento-de-navidad/28459149. (Original work published 1905)
- Dickens, C. (1891). *Rozhdestvenskaya pesen v proze* (S. M. Dolgova, Trans.). Runivers. (Original work published 1905)
- Dickens, C. (2021). *Rozhdestvenskaya pesen v proze* (T. Ozerskaya, Abr. ed.). ACT. (Original work published 1905)
- Dickens, C. (2004). *Un Chant de Noël* (L. Papineau, Abr. ed.). Héritage. (Original work published 1905)

COPRUS 3/3

- Twain, M. (1917). *Les Aventures de Tom Sawyer* (P. F. Caillé and Y. Dubois-Mauvais, Trans.). Ebooks libres et gratuits. (Original work published 1876)
- Twain, M. (1996). *Les Aventures de Tom Sawyer* (F. De Gaïl, Trans.). Flammarion. (Original work published 1876)
- Twain, M. (2003). *Las Aventuras de Tom Sawyer* (J. Torroba, Trans.). Biblioteca Virtual Universal. (Original work published 1876)
- Twain, M. (2007). *Las Aventuras de Tom Sawyer* (L. I. Barrena, Abr. ed.). Anaya. (Original work published 1876)
- Twain, M. (2010). *Las Aventuras de Tom Sawyer* (B. Palacios, Abr. ed.). Dirección General de Bibliotecas. (Original work published 1876)
- Twain, M. (1917). *Les Aventures de Tom Sawyer* (P. F. Caillé and Y. Dubois-Mauvais, Trans.). Ebooks libres et gratuits. (Original work published 1876)
- Twain, M. (1996). *Les Aventures de Tom Sawyer* (F. De Gaïl, Trans.). Flammarion. (Original work published 1876)
- Twain, M. (2020). *Les Aventures de Tom Sawyer* (A. Culleton, Abr. ed.). Broché. (Original work published 1876)
- Twain, M. (1972). *Priklyucheniya Toma Soyera* (K. Chukovskiy, Trans.). Kaliningradskoe Knizhnoe Izdatelstvo. (Original work published 1876)
- Twain, M. (2014). *Priklyucheniya Toma Soyera* (I. O. Rodin, Abr. ed.). Biblioteka Shkolnika. (Original work published 1876)
- Twain, M. (2014). *Priklyucheniya Toma Soyera* (N. L. Daruzes, Abr. ed.). Vita Nova. (Original work published 1876)
- Twain, M. (1876), *The Adventures of Tom Sawyer*. Project Gutenberg.
- Twain, M. *The Adventures of Tom Sawyer* (Abr. ed.). Global Publishing Solutions. https://americanenglish.state.gov/. (Original work published 1876)
- Twain, M. (2000). *The Adventures of Tom Sawyer* (J. Kehl, Trans.) Pearson Education. (Original work published 1876)
- Twain, M. (2000). *The Adventures of Tom Sawyer* (N. Bullard, Abr. ed.). Oxford University Press. (Original work published 1876)