The MultiplEYE Text Corpus: Towards a Diverse and Ever-Expanding Multilingual Text Corpus

Ramunė Kasperė[*1], Anna Bondar[*2], Sergiu Nisioi[*3], Maja Stegenwallner-Schütz[*4], Hanne B. Søndergaard Knudsen[5], Ana Matić[6], Eva Pavlinušić Vilus[2], Dorota Klimek-Jankowska[7], Chiara Tschirner[2], Not Battesta Soliva[2], Deborah N. Jakobi[2], Cui Ding[2], Dima Abu Romi[8], Cengiz Acarturk[9], Matilda Agdler[2,10], Anton Marius Alexandru[3], Mohd Faizan Ansari[11], Annalisa Arcidiacono[12], Elizabete Ausma Velta Barisa[13], Ana Bautista[14,15], Lisa Beinborn[16], Yevgeni Berzak[8], Nedeljka Bjelanović[17], Anna Isabelle Bothmann[18,2], Jan Brasser[2], Caterina Cacioli[19], Anila Çepani[20], Ilze Ceple[13], Adelina Çerpja[21], Dalí Chirino[22], Jan Chromý[23], Alessandro Corona Mendozza[10], Iria de-Dios-Flores[24], Nazik Dinçtopal Deniz[25], Ana Došen[6], Kristian Elersič[26], Inmaculada Fajardo[27], Zigmunds Freibergs[28,13], Angelina Ganebnaya[13], Shan Gao[2], Jéssica Gomes[29], Annjo Klungervik Greenall[30], Alba Haveriku[31], Miao He[32,2], Anamaria Hodivoianu[3], Yu-Yin Hsu[33], Amanda Isaksen[30], Andreia Janeiro[29], Kristine Jensen de López[5,34], Aleksandar Jevremovic[35], Vojislav Jovanović[36], Hanna Kędzierska[7], Nik Kharlamov[5], Sara Košutar[37], Nelda Kote[31], Vanja Kovic[36], Izabela Krejtz[38], Thyra Krosness[2,39], Oleksandra Kuvshynova[3], Eilam Lavy[40], Ella Lion[8], Marta Łockiewicz[41], Kaidi Lõo[28], Paula Luegi[29], Mircea Mihai Marin[3], Clara Martin[14,42], Svitlana Matvieieva[43], Diane C. Mézière[44], Xavier Mínguez-López[27], Valeriia Modina[45], Jurgita Motiejūnienė[1], Marie-Luise Müller[46], Tolgonai Nasipbek kyzy[47], Jamal Abdul Nasir[48], Johanne S. K. Nedergaard[10], Ayşegül Özkan[49], Patrizia Paggio[10], Marijan Palmović[6], Maria Christina Panagiotopoulou[2], Alberto Parola[10], Helena Pérez[50], Klaudia Petersen[10], Anja Podlesek[26], Eva Pospíšilová[23], Marta Prauliņa[13], Mikuláš Preininger[51], Loredana Pungă[52], Diego Rossini[47], Špela Rot[53], Habib Sani Yahaya[54], Irina A. Sekerina[45], Anne Gabija Skadiņa[13], Jordi Solé-Casals[55], Lonneke van der Plas[47], Saara M. Varjopuro[44], Spyridoula Varlokosta[56], João Veríssimo[29], Oskari Juhapekka Virtanen[44], Nemanja Vračar[57], Mila Vulchanova[30], Ahmad Mustapha Wali[3], Peizheng Wu[2], Nilgün Yücel[58], Stefan Frank[22], Nora Hollenstein[2], Lena A. Jäger[2]

[1]Kaunas University of Technology, Lithuania; [2]University of Zurich, Switzerland; [3]University of Bucharest, Romania; [4]University of Koblenz, Germany; [5]Aalborg University, Denmark; [6]University of Zagreb, Croatia; [7]University of Wrocław, Poland; [8]Technion – Israel Institute of Technology, Israel; [9]Jagiellonian University, Poland; [10]University of Copenhagen, Denmark; [11]Silesian University of Technology, Poland; [12]University of Bergen, Norway; [13]University of Latvia, Latvia; [14]Basque Center on Cognition, Brain and Language, Spain; [15]University of the Basque Country, Spain; [16]University of Goettingen, Germany; [17]Institute for Literature and Arts, Serbia; [18]University College London, UK; [19]University of Florence, Italy; [20]University of Tirana, Albania; [21]Academy of Sciences of Albania, Albania; [22]Radboud University, Netherlands; [23]Charles University, Czech Republic; [24]Universitat Pompeu Fabra, Spain; [25]Boğaziçi University, Türkiye; [26]University of Ljubljana, Slovenia; [27]University of Valencia, Spain; [28]University of Tartu, Estonia; [29]University of Lisbon, Portugal; [30]Norwegian University of Science & Technology, Norway; [31]Polytechnic University of Tirana, Albania; [32]University of Konstanz, Germany; [33]The Hong Kong Polytechnic University, Hong Kong; [34]Agder University, Norway; [35]Singidunum University, Serbia; [36]University of Belgrade, Serbia; [37]UiT The Arctic University of Norway, Norway; [38]SWPS University, Poland; [39]University of Applied Sciences Northwestern Switzerland, Switzerland; [40]The Hebrew University of Jerusalem, Israel; [41]University of Gdańsk, Poland; [42]Ikerbasque Basque Foundation for Science, Spain; [43]Dragomanov Ukrainian State University, Ukraine; [44]University of Turku, Finland; [45]City University of New York, USA;

[46]Leibniz Institute for Psychology, Germany; [47]Università della Svizzera italiana, Switzerland;
[48]University of Galway, Ireland; [49]Baskent University, Türkiye;
[50]University of Santiago de Compostela, Spain; [51]Czech Academy of Sciences, Czech Republic;
[52]West University of Timișoara, Romania; [53]St. Stanislav's Institution, Slovenia; [54]Gozak Media, Nigeria;
[55]University of Vic - Central University of Catalonia, Spain;
[56]National and Kapodistrian University of Athens, Greece; [57]University of Padua, Italy;
[58]Marmara University, Türkiye;

ramune.kaspere@ktu.lt, sergiu.nisioi@unibuc.ro, stegenwa@uni-koblenz.de, lenaann.jaeger@uzh.ch
multipleye.project@gmail.com

Abstract

We present the MultiplEYE Text Corpus, a large-scale, document-level, multi-parallel resource designed to advance cross-linguistic research on reading and language processing. The corpus provides paragraph-level alignment for texts in 39 languages spanning seven language families and seven scripts. Unlike many existing multilingual corpora, a substantial number of documents were originally written in languages other than English, reducing English-centric bias and supporting more typologically diverse investigations. The texts are carefully selected to balance linguistic richness with experimental feasibility, particularly for eye-tracking-while-reading studies. Developed within a multi-lab initiative, the MultiplEYE Text Corpus follows unified translation, alignment, and experimental design guidelines to ensure cross-linguistic comparability. Its inclusion of texts varying in type and difficulty enables research on discourse-level processing, genre effects, and individual differences across a wide range of languages. The text corpus and accompanying metadata provide a robust foundation for multilingual psycholinguistic and computational modeling research. Data and materials are publicly available at https://doi.org/10.23668/psycharchives.21750.

1. Introduction

Document-level, multi-parallel text corpora play an important role in advancing cross-lingual research in theoretical linguistics, natural language processing, and psycholinguistics (de Varda and Marelli, 2022; Hamilton and Huth, 2020; Läubli et al., 2018; Pal et al., 2024; Yang et al., 2022). In contrast to sentence-based corpora, they enable the investigation of discourse-level phenomena and the development of context-aware language technologies. In this paper, we present the MultiplEYE Text Corpus, a multi-parallel document-level resource that provides paragraph-aligned texts in 39 languages across seven language families and seven scripts. The text corpus is specifically designed for reading experiments using eye-tracking, but it can also support other behavioral methods, such as self-paced reading, and neurophysiological methods, including electroencephalography (EEG) in co-registration with eye-tracking. The texts are compact enough for use in experimental settings, yet sufficiently long and diverse to enable the study of discourse-level processing and comparisons across text types and languages. The corpus can also be used to investigate research questions outside the scope of behavioral and neurophysiological research. For example, the multi-parallel nature of the corpus makes it possible to conduct cross-linguistic research that goes beyond the comparison of individual language pairs. Moreover, the corpus coverage of both typologically distinct and closely related languages, together with varied scripts, supports research on language contact, cross-linguistic universals, and comparative analysis across different scripts. Unlike many existing datasets, a substantial part of the texts in the MultiplEYE Text Corpus were originally written in languages other than English, which helps reduce the English-centric bias that is common in linguistic resources. In addition, the corpus includes texts of varying type and difficulty; this diversity supports the investigation of a broad range of research questions that are often difficult to address because of resource scarcity. For example, recent research has demonstrated that text genre affects readers' eye movement patterns (Gómez-Merino et al., 2022) and interacts with established psycholinguistic phenomena, such as predictability effects (Bolliger and Jäger, 2025). The MultiplEYE Text Corpus enables detailed investigation of such questions across a wide range of languages.

The corpus was created within the multi-lab MultiplEYE European Cooperation in Science and Technology (COST) Action[1,2] aimed at building a large multilingual eye-tracking-while-reading dataset that supports cross-linguistic research in psycholinguis-

*Equal contribution

[1]https://multipleye.eu
[2]www.cost.eu/actions/CA21131/

tics and multilingual modeling (Jakobi et al., 2025). The MultiplEYE project provides translation and alignment guidelines (Hollenstein et al., 2026) and establishes the necessary infrastructure required for cross-linguistic comparability through a unified experimental design (covering stimulus selection and layout, procedure, and pre-processing), and shared FAIR (Findable, Accessible, Interoperable, and Reusable)-compliant (Wilkinson et al., 2016) resources for software, and (meta-) data management, storage, and sharing. The main outcome of the initiative will be a large, publicly available dataset of eye-tracking data collected across multiple European and non-European languages, with a special focus on the inclusion of low- and very low-resource languages.

In this paper, we present the multilingual text stimuli used in the MultiplEYE eye-tracking-while-reading experiment. We also document the text selection and translation procedures, describe cross-linguistic differences between the texts, and provide linguistic annotations to enable the use of the corpus beyond the scope of the MultiplEYE initiative.

2. Related Work

Multilingual corpora are essential for advancing both natural language processing and cross-linguistic behavioral research. However, despite the growing availability of corpora in Natural Language Processing (NLP) and psycholinguistics, most are optimized for a single research domain and are seldom suitable for computational modeling and controlled experimental use.

In recent years, numerous open-source multilingual corpora have been published. They span multiple languages and scripts and open new opportunities for cross-lingual and multilingual research. The growing interest in NLP and multilingual Large Language Models (LLMs) is reflected in the construction of these corpora: They are predominantly introduced within NLP- or machine-translation studies, where they primarily serve as resources for model training and evaluation. In addition, corpus coverage increasingly emphasizes low-resource languages and broader typological diversity, with recent datasets providing resources for low- and extremely low-resource languages (e.g., PARME, Ahmadi et al., 2025; Samanantar, Ramesh et al., 2022; WebCrawl African, Vegi et al., 2022). However, most recent of the open-source multilingual resources are not fully multi-parallel: they provide different texts or sentences for each language, typically aligned only with an English version, rather than with each other. In contrast, truly multi-parallel datasets, in which the same materials are available across all languages, are rare and tend to be smaller in size or limited to specific domains (e.g., four-language fully parallel dialogues such as XDailyDialog, Liu et al., 2023; or the OpenWHO corpus, Merx et al., 2025). Additionally, there has been a shift from sentence-level to paragraph- or document-level corpora, facilitating the development of context-aware systems. Regarding quality control, the data-driven nature of target applications has led to the creation of very large corpora, in which ensuring quality becomes increasingly challenging. While some corpora are human-translated and curated (e.g., FLORES-200, Gordeev et al., 2024; MASSIVE, FitzGerald et al., 2023; PARME, Ahmadi et al., 2025; OpenWHO, Merx et al., 2025; XDailyDialog, Liu et al., 2023), many others do not undergo unit-by-unit human review and instead rely on automatic quality evaluation (e.g., automatically checking the alignment between the English and target-language sentence or paragraph) (Schwenk et al., 2021b,a; El-Kishky et al., 2020). At the same time, several studies have shown that using high-quality data is beneficial for model training (Lee et al., 2022b; Wenzek et al., 2020). In terms of text types, most multi-parallel, text-level corpora contain texts from narrow domains or genres—e.g., parliamentary debates (Europarl, Koehn, 2005), official diplomatic documents (UN, Ziemski et al., 2016), scripted talks (WIT3, Cettolo et al., 2012), or film dialogues (OpenSubtitles2016, Lison and Tiedemann, 2016). At the same time, recent research has demonstrated that the diversity in training materials enhances model performance (Yu et al., 2022), underscoring the value of creating multi-parallel corpora that cover a broad range of text types.

Beyond datasets published in NLP, eye-tracking-while-reading studies have also released multilingual corpora. Psycholinguistic reading research is increasingly focusing on text-level stimuli, which allow the investigation of effects that go beyond single-sentence processing. One prominent example is MECO (The Multilingual Eye-Movement Corpus, Siegelman et al., 2022; Kuperman et al., 2023; Siegelman et al., 2025; Kuperman et al., 2025), which uses a curated set of short reading passages sourced from Wikipedia and matches across more than 30 languages to enable direct cross-linguistic comparison. The materials consist of expository prose (i.e., encyclopedic texts). For each language, texts were prepared based on English source texts: translators produced high-quality translations (or selected established translations where available), with native-speaker verification. Another notable example is GECO (the Ghent Eye-Tracking Corpus, Cop et al., 2017), which uses a novel-length narrative—Agatha Christie's The Mysterious Affair at Styles—to enable naturalistic reading analyses in English and Dutch with English monolinguals and Dutch-English bilinguals. The materials con-

sist of literary narrative fiction and are available as an open-access resource. Although both MECO and GECO provide text-level, multilingual/bilingual and multi-parallel materials, their limited text type and source language diversity restricts their generalizability.

In sum, there is a need for multilingual, multi-parallel, non-English-centric, human-curated text corpora that are both typologically diverse and vary in text types. The MultiplEYE Text Corpus responds to this need by providing resources for both high- and low-resource languages, encompassing multiple text types, and providing human-curated document-level alignment.

3. Text Corpus

3.1. Text Selection Procedure

Text selection was conducted by an international, multilingual team of researchers from the MultiplEYE European Cooperation in Science and Technology (COST) Action. First, the team compiled a candidate pool of texts that included multiple text types and were originally written in various languages. Second, the final set of texts was selected to ensure a balanced distribution of text types, diversity of original languages, and the availability of high-quality translations across most languages included in the corpus. When only a few paragraphs of a source text were included in the MultiplEYE Text Corpus, excerpts were chosen so that discourse coherence was preserved and each passage was self-contained, i.e., comprehensible without access to preceding material.

3.2. Description of Included Texts

The MultiplEYE Text Corpus comprises twelve texts. The texts cover the following text types: two popular-science texts, two institutional texts, five literary texts, two argumentative texts (both consisting of two parts), and one encyclopedic text (see Table 1 for an overview). The texts vary in register, difficulty, structure, proportion of domain-specific terms, and narration style. The following section provides an overview of the texts and their specific characteristics.

1. *The MultiplEYE Project* is a popular science expository text that describes the MultiplEYE COST Action eye-tracking data collection initiative. This is the only text written specifically for the MultiplEYE project. It was originally written in English by the members of the MultiplEYE Action and was subsequently revised by a native English-speaking writer specializing in popular science texts to ensure its clarity. It includes a wide range of domain-specific terminology (e.g., “eye-tracker”, “natural language processing”, “machine language processing”). The text presents factual information about the project and is designed to be accessible to a non-specialist audience.

2. *Swarthy, blue-eyed caveman revealed using DNA from ancient tooth* is a popular science expository/informative text on the archaeological findings of two Mesolithic hunter-gatherers, sourced from the newspaper *The Guardian*, written in English by Ian Sample in 2014. The lexis is mostly accessible, with a limited number of domain-specific terms (e.g., “DNA”, “Mesolithic”, “immune system”). The narrative mostly comprises paratactic sentences, especially at the beginning, with more hypotactic sentences in the main body.

3. *Universal Declaration of Human Rights – Preamble* is an institutional text, written by the UN General Assembly and adopted in 1948. The text is informative/expository and is structured as a series of parallel “Whereas” clauses with embedded clauses. Its lexis is predominantly abstract (e.g., “dignity”, “inalienable rights”) and nominalized (e.g., “protection”, “promotion”, “realisation”). The text also contains numerous named entities (e.g., “General Assembly”, “United Nations”, “Member States”).

4. *Progress report on a Learning Mobility Benchmark* is an EU institutional text, namely a progress report from the European Commission, as of 2017. It is an informative text, written in a formal, neutral, impersonal style, using standardized vocabulary and syntax, typical of EU institutional communication. It includes references to other legal acts and documents and contains numerous domain-specific terms (e.g., “learning mobility”, “employability”, “active citizenship”) and acronyms (e.g., “IVET”, “HE”). Sentences are mostly complex and predominantly hypotactic, with frequent subordinate purpose clauses.

5. *Broken April* is a literary text, originally written in Albanian by Ismail Kadare and published in 1978. It is narrated by a third-person narrator. The lexis is predominantly everyday, with culture-specific realia (e.g., “Kanun”, “kullas”) and toponyms (e.g., “Northern Plateau”, “Tirana”). The language is figurative, containing many metaphors, epithets, metonymies, etc. The text includes direct speech, and its syntax is a combination of complex hypotactic and short paratactic sentences.

6. *The Alchemist* is a literary text, originally written in Portuguese by Paulo Coelho and published in 1988. The story is told by a third-person narrator. The text is concise, and figurative language is infrequent but present (e.g., a simple metaphor: “as if some mysterious energy bound his life to that of

Text Type	Title	Orig. Lang.	Orig. Length	Eng. Length
Popular Science	The MultiplEYE Project	EN	811	811
	Swarthy, blue-eyed caveman revealed	EN	413	413
Institutional	Universal Declaration of Human Rights	–	327	327
	Progress report on a Learning Mobility Benchmark	–	531	531
Literary	The Alchemist – Chapter 1	PT	430	494
	The Magic Mountain – Foreword	DE	431	508
	Solaris – Chapter 2: The Solarists	PL	742	952
	Broken April – Chapter 3	SQ	610	600
Argumentative	The North Wind and the Sun	EL/EN	106	106
	Rapa Nui	EN/FR	703/797	703
	Cow's Milk	EN/FR	807/960	807
Encyclopedic	Wikipedia – The Moon	EN	103	103

Table 1: Texts included in the MultiplEYE Text Corpus, their text types, their original languages in which they were written (ISO 639 language code), and their approximate number of words in the original and English texts. The Universal Declaration of Human Rights and the Progress report on a Learning Mobility Benchmark were originally written in multiple languages.

the sheep"; a syllepsis: "thicker books [...] lasted longer, and made more comfortable pillows"). The syntax is predominantly simple and paratactic, with occasional temporal, conditional, and relative subordinate clauses.

7. *Magic Mountain* is a literary text, written by Thomas Mann in German and published in 1924. It contains many long, embedded sentences. Relative and complement clauses are frequently stacked within a single sentence, with frequent parenthetical insertions that serve a metadiscursive function and allow the author to make self-corrections via epanorthosis or commentary on the narration. The lexis contains figures of speech such as personification (e.g., "crisis shattered its way through life and consciousness") or metaphor (e.g., "story [...] covered with historic mould").

8. *Solaris* is a literary science-fiction text, written in Polish by Stanisław Lem and published in 1961. The lexis contains astrophysics terms (e.g., "Gamow-Shapley theory", "fluctuations of gravity") and neologisms related to astrophysics and biophysics (e.g., "planetophysicists", "solarists", "plasmic mechanism"). Most sentences are syntactically complex, containing relative and complement clauses.

9. *Rapa Nui* is an argumentative text originally written in English and French. It is sourced from the Programme for International Student Assessment (PISA) 2018, which evaluates the reading comprehension abilities of 15-year-olds (OECD, 2019). The text consists of two documents within a scenario in which readers imagine preparing for attending a talk on the topic. Following the PISA 2018 reading framework, this scenario simulates an educational situation. The first document is a first-person blog post that contains an embedded book review. The book review constitutes the second document and is presented as an argumentative summary, narrated in the third person. It contains headings, timestamps, and URLs introducing web-discourse fragments. It also contains named entities (e.g., "Rapa Nui", "Polynesians") and culture-specific terms (e.g., "moai"). Figurative language is limited. The blog employs simple and compound paratactic sentences describing sequential actions and includes temporal and relative clauses. The review shifts to longer, hypotactic sentences with complement that-clauses and relative clauses.

10. *Cow's Milk* is an argumentative text originally written in English and French, sourced from PISA 2018 (OECD, 2019). The text consists of two documents presented as web journal articles, set within a scenario in which students research the documents. The documents contain scientific argumentation and convey viewpoints, using biomedical terms (e.g., "cardiovascular health", "diabetes") and many named entities (e.g., "United States Department of Agriculture"). They cite external sources (e.g., "International Dairy Foods Association", "National Institutes of Health") and use slogans (e.g., "do a body good"). Links to websites and article headings are included. The texts feature long hypotactic sentences with complement that-clauses, relative clauses, and prepositional phrases.

11. *Wikipedia – The Moon* is an encyclopedic text, originally written in English. As an expository text, it contains domain-specific terms (e.g., "natural satellite", "orbital period", "tidal locking") and numerals/units (e.g., "384,400 km", "29.5 days", "1.28 ls"). The text employs both simple and complex sentences, featuring relative clauses and participial modifiers (e.g., "that equals its orbital period", "resulting from being tidally locked to Earth"). The

language is formal and scientific.

12. *The North Wind and the Sun* is a short Aesopian fable, originally composed in the 6th century BCE in Ancient Greek. As Aesop's fables were transmitted orally, the original wording is not available. The fable is instead preserved in later written traditions, including Ancient Greek and Latin text collections, which reflect adapted versions of the narrative. To ensure cross-linguistic consistency, we used an English version of the text as the reference for all the translations. The text in English was taken from Aesop Language Bank (Aesop Language Bank Team, 2010). The fable is a didactic narrative with anthropomorphized main characters ("The Wind" and "the Sun"). The lexis is concrete and accessible, i.e., it contains only high-frequency words, to be easily understood for a wide audience. The sentences are predominantly paratactic. The language is simple and concise, making the story easy to understand.

3.3. Paragraph Alignment

The texts are aligned across languages at the paragraph level. Because the corpus was primarily designed as stimuli for an eye-tracking-while-reading experiment, the texts were split across pages (also referred to as screens) for use in an eye-tracking-while-reading experiment. Because English is the only common language for all the members of the project, the English version served as the reference for alignment in all languages, except Swiss German, which used German as the reference. Screen splits were introduced primarily at the end of a paragraph. Long paragraphs in the English reference version sometimes required division across multiple screens. All translations were aligned as closely as possible to the English pagination. Any overflow text in the languages other than English was placed on an additional page, after which alignment with the English pagination was resumed. Where this was not feasible, an extra screen was inserted in order to preserve alignment with the English version for subsequent screens. Sentences were never split across screens. Minor cross-language misalignments at the page level occur in a small number of texts because we used a fixed presentation format with a pre-defined maximal number of lines per screen and characters per line across languages to ensure suitability for eye-tracking experiments. Specifically, for screens, language-specific characteristics such as greater average word-length or syntactically less compact constructions led to paragraphs being considerably longer than in English, leading to a shift in pagination in a few cases.

This alignment procedure not only ensures consistency across materials used in the eye-tracking-while-reading experiment, but also facilitates cross-linguistic comparisons. All deviations from the English version are documented in the stimulus deviation forms and the alignment form published with this corpus.

4. Language Selection

The corpus covers a broad set of language families and scripts (see Table 3 in Appendix), including high-resource, medium-resource and low-resource languages. Beyond Indo-European, the dataset includes languages from Afro-Asiatic, Uralic, Turkic, Inuit-Yupik-Unangan, and Sinitic language families, as well as the isolate language Basque. This typological diversity of languages allows for cross-family comparisons, as well as for comparisons between the branches or across languages of one family. The corpus also includes multiple writing systems. The majority of languages use the Latin script, complemented by Cyrillic, Greek, Hebrew, Arabic, Devanagari, and Chinese scripts. The corpus also includes a substantial number of low-resource languages, among which are Romansh, Swiss German, Albanian, Kalaallisut[3], Basque, and others. See Figure 1 for the full list of languages included in the corpus.

5. Balancing Consistency and Diversity

One of the principal challenges in building the corpus was balancing diversity and consistency. In terms of diversity, the goal was to maximize coverage across languages, scripts, original languages, and text types. On the consistency side, the challenges included ensuring uniform paragraph alignment across languages, applying standardized translation selection and translation/post-editing workflows, ensuring consistency in reporting deviations, and following the same procedures for identifying and resolving translation inconsistencies.

In order to ensure consistency while allowing for diversity, a native-speaking representative for each language (hereafter referred to as *language coordinator*) was invited to join the initiative. The language coordinators were in charge of coordinating the text preparation for their languages, and adding metadata. Detailed translation guidelines were provided to the language coordinators (Hollenstein et al., 2026), and regular meetings with both the language coordinators and translators were held to address inquiries and provide further clarification. Besides, regular working-group meetings were held

[3]We use the term Kalaallisut instead of Greenlandic because this term is more appropriate from a historical, sociopolitical, and linguistic perspective.

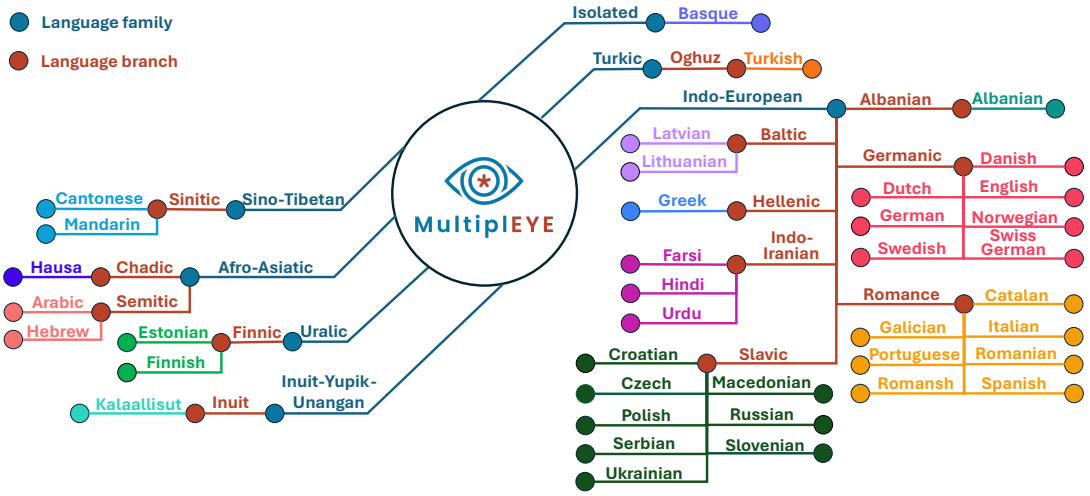

Figure 1: Languages represented in the MultiplEYE Text Corpus. Blue nodes indicate language families, orange nodes indicate branches, and the outer labels list the languages grouped under each branch.

to coordinate the creation of the multilingual text corpus, discuss ongoing translation, alignment, and documentation, share updates, and standardize procedures. Weekly virtual office hours were offered to support translation-related questions, corpus formatting (segmentation and alignment), and other inquiries, thus ensuring consistency across languages.

6. Translation

6.1. Translation Guidelines

Two pathways were distinguished depending on whether a published translation of the texts was available for a given language. When a published translation was available, the language coordinators were instructed to use it. Published translations were preferred because they are typically translated from the original language rather than from English, thereby supporting the reduction of the English-centering bias in the corpus. When multiple published translations existed, the most widely established version was preferred. Typographical, orthographic, or grammatical mistakes in available translations or inconsistent uses of punctuation (e.g., unsystematic use of single or double quotation marks) were corrected after verification by native speakers. However, stylistic changes were not permitted.

If no published translation of a given text was available, a new translation was produced by members of the MultiplEYE Action based on the English version of the text, except for Swiss German for which both English and German served as reference languages. The text *The MultiplEYE Project* was translated into all languages because it was written specifically for the MultiplEYE Text Corpus. The texts *Wikipedia – The Moon* and *Swarthy blue-eyed caveman revealed* were likewise translated to all of the languages included in the corpus by the members of the Action as no published translations for these specific texts were available.

To ensure comparability across languages, all translations followed a set of structured guidelines (Hollenstein et al., 2026). Translators were required to be native speakers of the target language with demonstrated proficiency in English. While professional translation experience was not mandatory, prior experience was considered advantageous. Translators were allowed to use machine translation systems (e.g., DeepL, DeepL GmbH); however, these outputs had to be revised manually and cross-checked for accuracy. It was recommended that translations be carried out independently and simultaneously by two persons. Back translation was recommended as a procedure for quality assurance. Besides, translators were encouraged to solve disagreements jointly. After that, all translations underwent manual revision by at least one additional native speaker to guarantee quality.

6.2. Translational Variation

For languages where no published translation was available, new translations were produced using the English text as the source, with the exception of Swiss German. For languages with existing published translations, these were carefully compared against the English reference, with observed deviations documented in the Stimulus Deviation Form (Nisioi et al., 2026).

In the following, we summarize the challenges encountered when preparing new translations as well as the differences observed in published trans-

lations.

For texts that required new translations, translators and language coordinators encountered several recurring challenges across languages in the areas of terminology, sentence structure, register and style, or orthography and punctuation.

Published translations differed across languages in terms of sentence structure and length (i.e., number of words), lexical and stylistic choices, and punctuation. When such differences from the English version were observed, they were documented. The published translations used for the corpus were not modified.

1. **Terminology gaps and lack of equivalents.** The most common challenge across many languages was the absence of established equivalents for domain-specific English terms (e.g., "eye-tracking", "COST Action", "learning mobility", "ECTS", "Europass", "active citizenship"). Translators reported having to choose between domestication (creating natural-sounding local terms) and foreignization (borrowing the English terms) strategies, also often applying explanatory, generalizing or descriptive paraphrasing transformations to render these terms accurately. Some lexical issues were unique to individual, most often low-resource, languages. For example, in Kalaallisut, research-related terms (e.g., "mind", "thought", "test") can be easily mistranslated into either terms connoting spirituality or medical terms connoting mental illness. Therefore, translators had to select context-appropriate equivalents to avoid unintended connotations. For Kalaallisut, some technical terms are already established and in use—often coined by the Language Secretariat of Greenland—such as *silassorissuusiaq* for "artificial intelligence" (roughly "a model/figurine of intelligence"). Other terms were coined by the translators as needed, for example *umiarsuit ulloriarsiuisussat* for "starships" (roughly "ships that will be out among the stars"). Given the polysynthetic morphology of Kalaallisut, the creation of novel words is common and such formations are generally natural for readers. Similarly, Romansh translations faced many incidental lexical gaps. However, these differed from those encountered in other languages. Although the relevant terminology exists, it is often very low-frequency and can therefore be confusing even for the native speakers. This raised the question of whether a technically accurate but unnatural translation should have been preferred over a vaguer formulation, sometimes requiring substantial circumlocution. To keep the difficulty of the text comparable, highly unusual words were avoided. As a side effect of the large number of words that are no longer in common use, Romansh speakers are aware of the existence of such low-frequency expressions and of gaps in their own knowledge. This can lead them to assume that unfamiliar words are highly specific or outdated Romansh vocabulary, whereas they are in fact foreign words. A common practice in written Romansh is therefore to mark foreign words in italics or with quotation marks. We followed this practice for foreign words in this corpus (e.g. "moai", "kulla"), which could otherwise be misinterpreted as unfamiliar Romansh vocabulary.

2. **Differences in Content.** Another challenge faced by the translators and language coordinators was variation in the content of the fable *The North Wind and the Sun*. Where multiple versions existed in a given language, the version closest to the English source was selected.

3. **Wordplay.** The title of the text *The MultiplEYE Project* posed problems because of the English pun which reflects the project's focus on multilingualism and eye-tracking research. In most languages, the translations cannot replicate this double meaning in a single word. Therefore, translators were recommended to provide explanations, most often in brackets, to preserve the meaning of the wordplay.

4. **Differences in sentence structure.** In newly produced translations frequent syntactic differences arose between English and the target languages. For example, translators had to reconcile the long, complex sentences typical of English with the preference of shorter, clearer structures in Mandarin, balancing fidelity to the source text with naturalness and readability. In published translations, sentence boundaries did not always correspond one-to-one across languages. In some cases, two English sentences corresponded to a single sentence in the target language (or vice versa), as observed for example in Croatian, Dutch, Lithuanian, and German.

5. **Register and style choices.** Translations into some languages required a careful selection of register and style. This was particularly relevant for languages spoken in bilingual regions, where the other language is dominant in most written communication. For example, Cantonese lacks a single standard written form. Given that the experiment for which the corpus was primarily designed involves a reading task rather than colloquial spoken interaction, a more neutral register in the Cantonese translations was adopted to ensure stylistic consistency across the different genres included in the materials. While it is true that spoken Cantonese lacks a standardized script system, Standard Written Chinese as used in Hong Kong (standard Hong Kong Cantonese) does follow established conventions and is the norm in formal written contexts such as education, news media, and government publications. This standard form draws

on traditional Chinese characters and largely follows Mandarin-based grammar, though it still exhibits several lexical and grammatical differences from Mandarin. The translation choice reflected this register, while also considering what is widely recognized and read by Cantonese speakers in Hong Kong as part of their everyday literate experience. Similar challenges were encountered for Romansh. In Romansh-speaking regions, German is often the only language present for many written usages. This means that the variety in registers for written Romansh can be more limited and the line between written and spoken Romansh is blurrier, so that some texts follow a more oral ductus in order to still be perceived as authentic usage of the language. At the same time there are tendencies in spoken Romansh that are traditionally strongly avoided in written language. The normalized integrations of German vocabulary into the spoken language is considered very bad practice in written form, so a balance had to be found between authentic tone and stylistically apt formulations.

6. **Orthographic and punctuation adjustments.** Minor but necessary adaptations were made to capitalization in languages such as Croatian, Danish, Latvian, and Lithuanian, where, for example, a capital letter used in the English version would have introduced an error in the target text. In such instances, lowercase was used to comply with the orthographic conventions of the target language. Punctuation was also adjusted in Croatian, Romanian, and other languages to comply with orthographic conventions. For a few languages, archaic spellings in published translations were updated to contemporary standards where they could occur unusual to the readers or hinder comprehension, while preserving the original tone and register. For Swiss German that is primarily oral and lacks standardized orthographic conventions, translators had to establish new spelling conventions and balance consistency with readability, whenever these conventions produced forms that might feel unnatural to readers.

7. **Language-specific structural challenges.** In some low-resource languages (e.g., Swiss German), the absence of a tense corresponding to the past perfect required translators to convey a past-perfect meaning using adverbial constructions in order to preserve temporal sequencing. Similarly, because Swiss German lacks a genitive case or a dedicated possessive marker, possession is expressed using a preposition–dative construction. This can result in multi-noun possessive chains and often required additional rephrasing to maintain readability.

In general, the differences between languages created challenges that necessitated targeted adaptations while maintaining as much fidelity to the source versions as possible.

7. Corpus Statistics

The corpus is accompanied by linguistic annotations and summary statistics to support cross-lingual analysis. For each language and text, we provide automatic tokenization,[4] and, where available, automatic sentence segmentation, part-of-speech tags, lemmas, and frequency counts.

The annotations are generated using an automated pipeline based on the spaCy large models (Montani et al., 2023)[5] with several exceptions.[6] For Kalaallisut and Hausa, we employ the multilingual small model `xx_sent_ud_sm`, which performs basic tokenization and rule-based sentence segmentation, while for Romansh we use Italian models. Several languages are not natively supported in spaCy; therefore, for Turkish and Hebrew we rely on community spaCy pipelines (Altinok, 2023; Zeldes et al., 2022). Farsi is analyzed using `hazm` (Roshan Research, 2023), Cantonese using the PyCantonese package (Lee et al., 2022a), and Hindi using the IndicNLP toolkit (Kunchukuttan, 2020).

We conduct cross-linguistic comparisons at the page-level, as pages represent the fundamental presentation units in user-facing eye-tracking experiments. Page-level statistics are directly comparable across languages and help mitigate differences in text length (Koplenig, 2019).

Table 2 reports for each language the number of tokens, the type-token ratio, the total number of sentences, the average word length, and the average number of words per page. The differences in these metrics primarily reflect typological properties. The number of tokens varies moderately across the corpus, with most languages containing between roughly 5,000 and 7,000 tokens. Some languages fall outside this range: Finnish and Estonian have comparatively smaller token counts, while Urdu and Hindi contain larger numbers of tokens.

With respect to the number of words per page, agglutinative and morphologically complex languages, such as Estonian, Finnish, Lithuanian, Turkish, Kalaallisut, and Basque, tend to have fewer words and longer average word lengths. In contrast, Germanic languages exhibit between 56 and 64 words per page, followed by Romance languages, Greek, Albanian, and Hausa, which display higher counts

[4] A token is the output result of a tokenizer, it can mean a word, an `<eos>` marker, a punctuation symbol, or a line break.

[5] spaCy version 3.8.7

[6] https://github.com/senisioi/multipleye_text_processing

Language	#Wds	TTR	#Snts	Wd. Len.	#Wds pp
English	6405	29%	287	4.79	64
German	6013	37%	327	5.89	56
Swiss German	6147	35%	827	5.24	57
Dutch	6511	31%	343	5.22	63
Danish	5924	35%	317	5.19	59
Swedish	5813	37%	322	5.27	58
Norwegian	5806	35%	351	5.11	58
Basque	5089	49%	314	6.71	40
Catalan	6937	29%	289	4.63	69
Spanish	6938	31%	297	4.98	69
Portuguese	6643	32%	310	4.92	65
Italian	6321	35%	333	5.24	61
Romansh	6862	32%	300	5.05	67
Romanian	6317	38%	299	5.18	63
Galician	6618	34%	291	5.02	66
Slovenian	5654	45%	301	5.25	55
Croatian	5538	47%	298	5.48	54
Serbian	5634	47%	290	5.27	56
Macedonian	6071	39%	497	5.28	60
Polish	5431	50%	296	6.01	54
Czech	5305	50%	296	5.43	53
Ukrainian	5294	49%	297	5.73	52
Russian	5295	49%	287	6.01	52
Lithuanian	4920	54%	310	6.28	49
Latvian	5089	50%	319	5.95	50
Estonian	4605	56%	306	6.55	46
Finnish	4356	60%	307	7.49	42
Albanian	6747	34%	288	4.82	66
Greek	6598	37%	327	5.45	65
Turkish	4754	55%	330	6.50	47
Arabic	4953	54%	261	4.79	50
Hebrew	5002	58%	293	4.63	50
Hausa	7222	21%	284	4.43	70
Hindi	7115	27%	103	4.04	69
Urdu	7609	22%	292	3.51	76
Farsi	5556	31%	100	3.82	55
Mandarin	6140	34%	284	1.74	60
Cantonese	5870	36%	101	1.75	58
Kalaallisut	3332	71%	283	12.15	31

Table 2: Statistics by language: the number of tokens for each language (#Wds), the type-token ratio (TTR), the total number of sentences (#Snts), the average word length (Wd. Len.), and the average number of words per page (#Wds. pp). Kalaallisut is a polysynthetic language and type-token ratio is not a meaningful measure. Color-codes represent language sub-families.

(ranging from 61 to 70 words per page). Texts written in languages that use abjads, such as Hebrew and Arabic, show similar numbers of words per page, as do texts in logographic languages such as Mandarin and Cantonese. Hindi is the only language in the corpus written in an abugida.

A notable exception is Kalaallisut, with an average of 31 words per page, reflecting the polysynthetic nature of the language. A single Kalaallisut word often corresponds to an entire sentence in a language such as English.

With respect to the type–token ratio, three trends can be observed. Languages with a high type–token ratio include Slavic and Baltic languages (e.g., Polish, Russian, Ukrainian, Czech, and Lithuanian), as well as Turkish, Arabic, Hebrew, Basque, Finnish, Estonian, and Kalaallisut, the latter exhibiting the highest ratio. A middle band comprises the Romance languages (French, Italian, Portuguese, and Romanian) and several Germanic languages, with English and Dutch at the lower end of this group. The lowest ratios are observed for Hausa, Urdu, and Hindi.

Taken together, these metrics can be used to reconstruct an approximate phylogenetic tree of languages (see Figure 2 in the Appendix) using Ward's method over the Euclidean distance matrix computed from the average number of unique words per page, the percentage of function words, and the average word length.

8. Accessing Data and Metadata

Texts in the corpus are accompanied by metadata detailing bibliographic and provenance information for both the original and translated versions to ensure transparency, traceability, and filtering within the corpus. The text corpus, annotations and accompanying metadata have been publicly released through an open-access online repository (Nisioi et al., 2026). Due to copyright restrictions, only the linguistic annotations and the metadata are available for literary texts and their translations. Newly created translations of copyright free materials may be used for research purposes. For inquiries about specific language versions, contact details of the respective language coordinator are available via the repository.

9. Conclusion

The MultiplEYE Text Corpus comprises human-curated, multi-way parallel, document-level, paragraph-aligned annotated textual materials for both high- and low-resource languages from typologically different language families, covering diverse text types. Additionally, it provides page-level aligned materials, particularly useful for eye-tracking-while-reading studies.

10. Contribute

MultiplEYE is an open initiative inviting researchers to join the Action and contribute additional languages following shared guidelines on translation and post-editing workflow, deviation reporting, and metadata, so that the texts remain comparable across the corpus. For more information about the project and possible contributions please see Jakobi et al. (2025).

11. Acknowledgments

This work was supported by the European Cooperation in Science and Technology (COST) under COST Action CA21131 (MultiplEYE), the Basque Government (grant: BERC 2022–2025 program), the Carlsberg Foundation (grants: CF24-2005, CF23-1627), the Croatian Science Foundation (grant: IPCH-2022-04-3316), the Czech Science Foundation (grant: 23-06796S), the Department of Language & Literature, Norwegian University of Science & Technology (Strategic Funding grant, 2024), the Digital Society Initiative at the University of Zurich (PhD-scholarship), the Dutch National Science Organisation (grant: VI.Veni.211C.039), the Foundation for Research in Science and the Humanities at the University of Zurich (grant: Creating a Multilingual Eye-Tracking Corpus . . .), the Foundation for Science and Technology, Portugal (grants: UID/214/2025, UI/BD/154500/2022), the Estonian Research Council (grant: PSG743), the French National Research Agency (grant: ANR-24-MRS1-0003), the German Federal Ministry for Education and Research (grant: 01IS20043), the Hong Kong Polytechnic University-Areas of Excellence (grant: P0045115), the Independent Research Fund Denmark (grant: 2027-00079B), the Israel Science Foundation (grant: 1499/22), the Jagiellonian University Strategic Programme Excellence Initiative (grant: ID.UJ), the Kadri, Nikolai and Gerda Rõuk Research Fund at the University of Tartu, the "la Caixa" Foundation (grant: LCF/BQ/DR23/12000006), the Latvian Council of Science (grant: lzp-2025/1-0195), the Romanian National Research Council (grant: PN-IV-P2-2.1-TE-2023-2007), the Slovenian Research and Innovation Agency (grant: P5-0110), the Spanish State Research Agency (grants: BCBL Severo Ochoa excellence accreditation CEX2020-001010/AEI/10.13039/501100011033; PID2023-148756NB-I00), the Swiss National Science Foundation (grants: 212276, 10002551, 225146), swissuniversities (grants: OpenEye, EyeStore, EyeStore+), the SWPS University (grant: 52/2025/FRBN/G), and zukunft.niedersachsen (Impulsprofessur grant).

References

Aesop Language Bank Team. 2010. Aesop Language Bank.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin Hamaamin, Saya Hama, Ardeshir Mousavi, Sarko Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Ramezani, Aryan Aziz, Ali Salehi Sheikhalikelayeh, Mohammadreza Yadegari, Kewyar Yadegari, and Sedighe Roodsari. 2025. PARME: Parallel corpora for low-resourced Middle Eastern languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 30032–30053. Association for Computational Linguistics.

Duygu Altinok. 2023. A diverse set of freely available linguistic resources for Turkish. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13739–13750. Association for Computational Linguistics.

Lena Bolliger and Lena A. Jäger. 2025. Genre matters: How text types interact with decoding strategies and lexical predictors in shaping reading behavior. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7459–7476. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268. European Association for Machine Translation.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics*, pages 138–144. Association for Computational Linguistics.

DeepL GmbH. DeepL translator. https://www.deepl.com/en/translator.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document Pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez,

Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4277–4302. Association for Computational Linguistics.

Nadina Gómez-Merino, Inmaculada Fajardo, Antonio Ferrer, and Holly Joseph. 2022. Eye movements of deaf students in expository versus narrative texts. *American Annals of the Deaf*, 167(3):313–333.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. FLORES+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623. Association for Computational Linguistics.

Liberty Hamilton and Alexander Huth. 2020. The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582.

Nora Hollenstein, Marie-Luise Müller, Deborah N. Jakobi, Cui Ding, Maja Stegenwallner-Schütz, Ana Matić, Eva Pavlinušić Vilus, Ramunė Kasperė, Anna Bondar, Maroš Filip, Stefan Frank, Jana Hofmann, Thyra Krosness, Kaidi Lõo, Johanne Nedergaard, Chiara Tschirner, and Lena A. Jäger. 2026. *MultiplEYE Data Collection Guidelines*.

Deborah Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine Jensen de López, Nik Kharlamov, Hanne Søndergaard Knudsen, Yevgeni Berzak, Ella Lion, Irina Sekerina, Cengiz Acarturk, Mohd Ansari, Katarzyna Harezlak, Pawel Kasprowski, Ana Bautista, Lisa Beinborn, Anna Bondar, Antonia Boznou, Leah Bradshaw, Jana Hofmann, Thyra Krosness, Not Soliva, Anila Çepani, Kristina Cergol, Ana Došen, Marijan Palmovic, Adelina Çerpja, Dalí Chirino, Jan Chromý, Vera Demberg, Iza Škrjanec, Nazik Deniz, Inmaculada Fajardo, Mariola Giménez-Salvador, Xavier Mínguez-López, Maroš Filip, Zigmunds Freibergs, Jéssica Gomes, Andreia Janeiro, Paula Luegi, João Veríssimo, Sasho Gramatikov, Jana Hasenäcker, Alba Haveriku, Nelda Kote, Muhammad Kamal, Hanna Kedzierska, Dorota Klimek-Jankowska, Sara Kosutar, Daniel Krakowczyk, Izabela Krejtz, Marta Łockiewicz, Kaidi Lõo, Jurgita Motiejūnienė, Jamal Nasir, Johanne Nedergård, Ayşegül Özkan, Mikuláš Preininger, Loredana Pungă, David Reich, Chiara Tschirner, Špela Rot, Andreas Säuberli, Jordi Solé-Casals, Ekaterina Strati, Igor Svoboda, Evis Trandafili, Spyridoula Varlokosta, Mila Vulchanova, and Lena Jäger. 2025. MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*. Association for Computing Machinery.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Alexander Koplenig. 2019. A non-parametric significance test to compare corpora. *PLoS One*, 14(9):e0222703.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Victor Kuperman, Sascha Schroeder, Cengiz Acartürk, Niket Agrawal, Dominick M. Alexandre, Lena S. Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Luiz Vinicius Gadelha De Freitas, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, Charlotte E. Lee, Jun Ren Lee, Marina P. T. Leite, Simona Mancini, Nataša Mihajlović, Ksenija Mišić, Miloslava Orekhova, Olga Parshina, Milica Popović Stijačić, Athanassios Protopapas, David R. Reich, Anurag Rimzhim, Rui Rothe-Neves, Thais M. M. Sá, Andrea Santana Covarrubias, Irina Sekerina, Heida M. Sigurdardottir, Anna Smirnova, Priyanka Srivastava, Elisangela N. Teixeira, Ivana Ugrinic, Kerem Alp Usal, Karolina Vakulya, João M. M. Vieira, Ark Verma, Denise H. Wu, Jin Xue, Sunčica Zdravković, Junjing Zhuo, Laoura Ziaka, and Noam Siegelman. 2025. New data on text reading in English as a second language: The Wave 2 expansion of the Multilingual Eye-Movement Corpus (MECO). *Studies in Second Language Acquisition*, 47(2):677–695.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Kaidi Lõo, Marco Marelli, Kelly Nisbet, Timothy C. Papadopoulos, Athanassios Protopapas, Satu Savo, Diego E. Shalom, Natalia Slioussar, Roni Stein, Longjiao Sui, Analí

Taboh, Veronica Tønnesen, and Kerem Alp Usal. 2023. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 45(1):3–37.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796. Association for Computational Linguistics.

Jackson L. Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022a. PyCantonese: Cantonese linguistics and NLP in python. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6607–6611. European Language Resources Association.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8424–8445. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 923–929. European Language Resources Association.

Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023. XDailyDialog: A multilingual parallel dialogue corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12240–12253. Association for Computational Linguistics.

Raphaël Merx, Hanna Suominen, Trevor Cohn, and Ekaterina Vylomova. 2025. OpenWHO: A document-level parallel corpus for health translation in low-resource languages.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Sergiu Nisioi, Anna Bondar, Ramuné Kasperé, and Maja Stegenwallner-Schütz. 2026. The multipleye text corpus data and materials.

OECD. 2019. *PISA 2018 Assessment and Analytical Framework*. OECD Publishing, Paris, France.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13185–13197. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Roshan Research. 2023. Hazm: Persian natural language processing toolkit. Python library for processing Persian text including normalization, tokenization, lemmatization, POS tagging, and dependency parsing.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1351–1361. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6490–6500. Association for Computational Linguistics.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, Marco Marelli, Timothy Papadopoulos, Athanassios Protopapas, Satu Savo, Diego Shalom, Natalia Slioussar, Roni Stein, Longjiao Sui, Analí Taboh, Veronica Tønnesen, Kerem Usal, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S. Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, Charlotte E. Lee, Jun Ren Lee, Marina P. T. Leite, Simona Mancini, Nataša Mihajlović, Ksenija Mišić, Miloslava Orekhova, Olga Parshina, Milica Popović Stijačić, Athanassios Protopapas, David R. Reich, Anurag Rimzhim, Rui Rothe-Neves, Thais M. M. Sá, Andrea Santana-Covarrubias, Irina Sekerina, Heida M. Sigurdardottir, Anna Smirnova, Priyanka Srivastava, Elisangela N. Teixeira, Ivana Ugrinic, Kerem Alp Usal, Karolina Vakulya, Ark Verma, João M. M. Vieira, Denise H. Wu, Jin Xue, Sunčica Zdravković, Junjing Zhuo, Laoura Ziaka, and Victor Kuperman. 2025. Wave 2 of the Multilingual Eye-Movement Corpus (MECO): New text reading data across languages. *Scientific Data*, 12:1183.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R, and Chitra Viswanathan. 2022. WebCrawl African: A multilingual parallel corpora for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012. European Language Resources Association.

Mark D Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J Lusher, Maryann E Martone, Albert Mons, Abel L Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Jibiao Yang, Antal van den Bosch, and Stefan Frank. 2022. Unsupervised text segmentation predicts eye fixations during reading. *Frontiers in Artificial Intelligence*, 5:731615.

Yu Yu, Shahram Khadivi, and Jia Xu. 2022. Can data diversity enhance learning generalization? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945. International Committee on Computational Linguistics.

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4331–4344. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3530–3534. European Language Resources Association.

Appendix

Code	Language	Script	Family	Branch
ar	Arabic	Arabic	Afro-Asiatic	Semitic
he	Hebrew	Hebrew	Afro-Asiatic	Semitic
ha	Hausa	Latin	Afro-Asiatic	Chadic
eu	Basque	Latin	Isolate	–
kl	Kalaallisut	Latin	Inuit-Yupik-Unangan	Inuit
tr	Turkish	Latin	Turkic	Oghuz
et	Estonian	Latin	Uralic	Finnic
fi	Finnish	Latin	Uralic	Finnic
yue	Cantonese	Traditional Chinese	Sino-Tibetan	Sinitic
zh	Mandarin	Simplified Chinese	Sino-Tibetan	Sinitic
lv	Latvian	Latin	Indo-European	Baltic
lt	Lithuanian	Latin	Indo-European	Baltic
el	Greek	Greek	Indo-European	Hellenic
ro	Romanian	Latin	Indo-European	Romance
gl	Galician	Latin	Indo-European	Romance
it	Italian	Latin	Indo-European	Romance
es	Spanish	Latin	Indo-European	Romance
ca	Catalan	Latin	Indo-European	Romance
pt	Portuguese	Latin	Indo-European	Romance
rm	Romansh	Latin	Indo-European	Romance
sq	Albanian	Latin	Indo-European	Albanian
hi	Hindi	Devanagari	Indo-European	Indo-Iranian
ur	Urdu	Arabic	Indo-European	Indo-Iranian
fa	Persian/Farsi	Arabic	Indo-European	Indo-Iranian
cz	Czech	Latin	Indo-European	Slavic
pl	Polish	Latin	Indo-European	Slavic
sr	Serbian	Cyrillic/Latin	Indo-European	Slavic
hr	Croatian	Latin	Indo-European	Slavic
sl	Slovenian	Latin	Indo-European	Slavic
uk	Ukrainian	Cyrillic	Indo-European	Slavic
mk	Macedonian	Cyrillic	Indo-European	Slavic
ru	Russian	Cyrillic	Indo-European	Slavic
no	Norwegian	Latin	Indo-European	Germanic
de	German	Latin	Indo-European	Germanic
zd	Swiss German	Latin	Indo-European	Germanic
sv	Swedish	Latin	Indo-European	Germanic
en	English	Latin	Indo-European	Germanic
nl	Dutch	Latin	Indo-European	Germanic
da	Danish	Latin	Indo-European	Germanic

Table 3: Languages in the corpus, with script, family, and branch

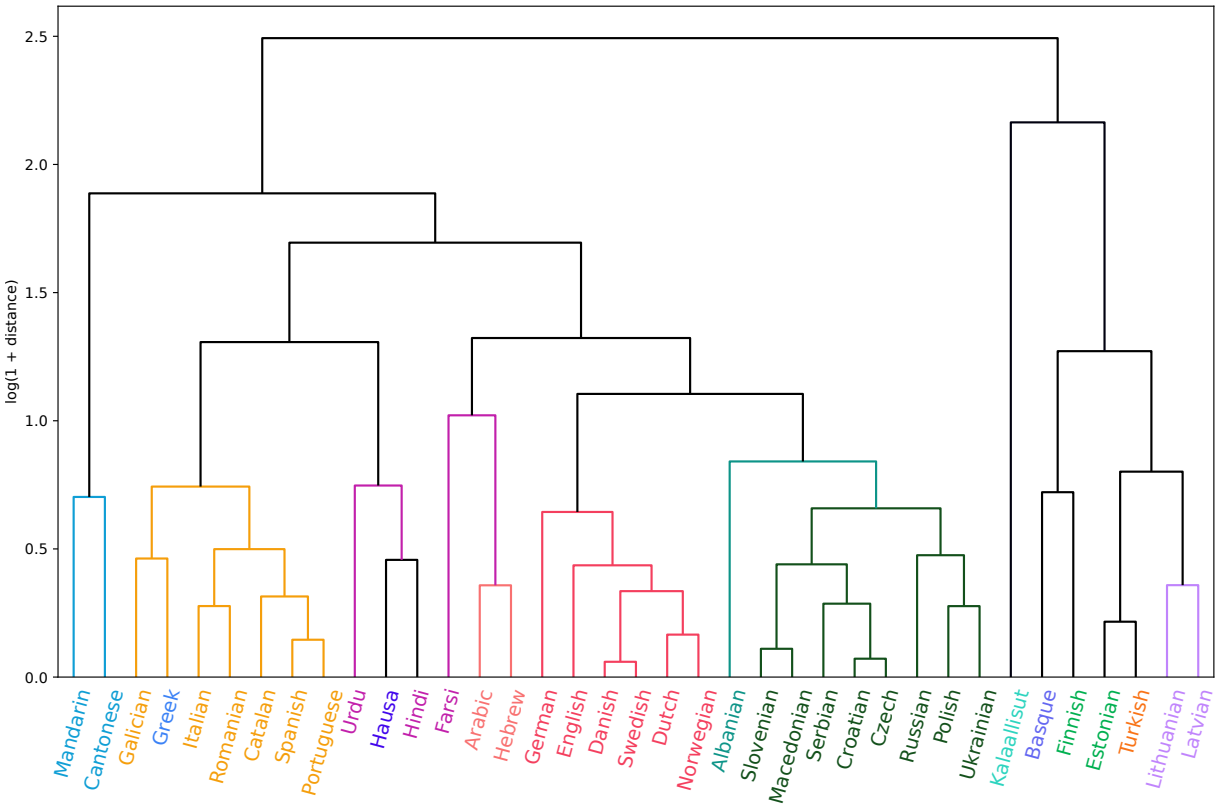

Figure 2: Clustering result representing an approximate reconstruction of the phylogenetic tree of languages. Several language families are well clustered together using only a small set of features measured at the page-level: average number of unique words, the percentage of function word, and word length. Ward's method was used for clustering. The y-axis represents the log of the Euclidean distance.