information

MDPI

Article

Translation Alignment with UGARIT

Tariq Yousef [1,*], Chiara Palladino [2], Farnoosh Shamsian [3] and Maryam Foradi [4]

1 Department of Computer Science, Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany
2 Department of Classics, Furman University, 3300 Poinsett Highway, Greenville, SC 29613, USA; chiara.palladino@furman.edu
3 Department of Ancient History, Leipzig University, Beethovenstraße 15, 04107 Leipzig, Germany; farnoosh.shamsian@uni-leipzig.de
4 Institute of Applied Linguistics and Translation Studies, Leipzig University, Beethovenstraße 15, 04107 Leipzig, Germany; maryam.foradi@uni-leipzig.de
* Correspondence: tariq.yosef@uni-leipzig.de

Abstract: UGARIT is a public web-based tool for manual annotation of parallel texts for generating word-level translation alignment. We aimed to develop a user-friendly interactive interface to visualize aligned texts and collect training data in the form of translation pairs to be used later, (i) for training an automatic translation alignment system for historical languages at the word/phrase level, (ii) as a gold standard to evaluate automatic alignment and machine translation systems. UGARIT is now widely used for learning new languages, especially historical languages, and as a reading environment for parallel texts. In the following sections, we present the related works and similar projects; then, we give an overview of the visualization techniques used to present the alignment results. Further, we explain how we could derive the translation graph from the aligned translation pairs. Finally, we discuss the usage limitations of UGARIT, possible improvements, and future development plans.

Keywords: translation alignment; text visualization; manual alignment; translations graph; human–computer interaction; user-centered design

Citation: Yousef, T.; Palladino, C.; Shamsian, F.; Foradi, M. Translation Alignment with UGARIT. *Information* **2022**, *13*, 65. https://doi.org/10.3390/info13020065

Academic Editor: Willy Susilo

Received: 5 November 2021
Accepted: 25 January 2022
Published: 27 January 2022

1. Introduction

Translation alignment is a major task in Digital Humanities and Natural Language Processing. It is the process of comparing two texts in different languages to find translation correspondences among the textual units in the source and translation texts [1]. It can be performed at various granularity levels according to the project's context or the research purpose.

Translation alignment is essential in neural and statistical machine translation [2], cross-lingual annotation projection [3,4], and translation lexica induction [5–7]. Several automatic approaches have been developed to perform the alignment at different levels [1,8–10]. However, most proposed models employ unsupervised statistical methods to generate alignment probabilities distribution between the source and target textual units. In general, The accuracy of the automatic alignment varies according to multiple factors, such as text type and length, size of the corpus, and translation quality and consistency. Recently, with the advances in transformers and contextualized language models [11,12], researchers have developed many approaches that can exploit contextualized multilingual word embeddings to generate word alignments from parallel texts [13,14].

In turn, manual alignment is still essential despite the advances in automatic translation alignments models, especially for creating alignment gold standards [15–19] and automatic alignment post-editing. However, manual alignment is expensive in terms of time and resources and requires annotation tools tailored for this purpose. Most word alignment annotation tools are either crafted for a specific project [20,21], specific language pairs, or specific texts [20,22].

The related works can be grouped into two main groups; the first one includes tools that offer an annotation interface to generate translation equivalents to be used in further development without any interest in visualizing the alignment. The second group includes tools that provide both the annotation and the visualization interfaces.

From the first group, we can mention the *Blinker Project* [20], which developed the first annotation tool for manual text alignment to align different versions of the Bible in French and English at the word level. The Linguistic Data Consortium (LDC) has also developed the *LDC Word Aligner* to perform the manual word alignment for Arabic–English and Chinese–English parallel texts from newswire and broadcast mainly [21]. Furthermore, *TagAlign* [23] allows users to annotate bilingual texts with a pre-defined tagset and create manual alignments at the sentence level. The tool was initially developed to align Brazilian, Portuguese, and English parallel texts. D. Benner [22] developed a manual word alignment tool to create gold standard for Hebrew–English by aligning the Hebrew bible with its English translation.

On the other hand, *Yawat* [24], *Alpheios* [25], *SWIFT Aligner* [26], and *CLUE-Aligner* [27], that fall into the second group, enable users to create their alignments manually at the word- and phrase-level and offer various possibilities for visualizing the aligned texts, such as side-by-side view, the interlinear text view [28], and alignment matrices [24,27]. UGARIT falls in the second group, it provides a manual annotation service and a reading environment for parallel texts.

The tools mentioned above use different approaches to link words in the source text with their correspondents in the translation. We can distinguish three main methods:

(i) The two texts are placed on two parallel columns. Annotators can draw lines between corresponding words of the source and translation texts [22,26];
(ii) The two texts are represented as a two-dimensional alignment matrix. Annotators can select a cell in the grid with a mouse click to create an alignment between the corresponding row- and column-words [27,29];
(iii) The texts are placed side-by-side. Annotators can select the source words and their translation correspondents with mouse clicks.

The first two approaches are suitable for short text units (sentences, short paragraphs), where annotators can still capture the context while reading vertical texts, whereas the third approach is suitable for both short and long texts. Most tools mentioned above are limited to one alignment class between tokens in the original and its translation, *CLUE-Aligner* [27] distinguishes between two main classes, possible and sure alignments and *Yawat* [24] supports manual labelling of the alignment relations. *SWIFT Aligner* [26] provides support for Part-Of-Speech and syntactic dependency manual annotation.

Visualization of aligned texts was the subject of interest and research in recent years; many tools have been developed for this purpose. Various approaches have been utilized to visualize the text alignment at different levels [30,31], for instance, side-by-side views, parallel views, and text heat maps.

UGARIT is a crowd-sourcing project that enables users to create translation alignments at the word or phrase level; the resulting translation pairs can be used as gold standards to evaluate machine translation systems or create dynamic lexica and translation memories. UGARIT was initially designed to visualize the automatically aligned texts available at *Perseus Digital Library* [32] and collect training data in the form of translation pairs to implement a statistical translation alignment system for historical languages, mainly Ancient Greek, Latin, and Persian, for which few to none aligned datasets exist. Ideally, historical languages are closed systems with a finite number of words and minimal change in the foreseeable future. Therefore, it should be possible to create adequately efficient automated alignment methods based on a relatively small training dataset. UGARIT is not only an annotation tool, but it also offers a visually powerful reading environment, where the reader can analytically compare texts token by token and at the same time observe the results through interactive visualization and statistics. Unlike other manual alignment tools, UGARIT collects the translation equivalents created by annotators to construct the trans-

lations graph, which can be used for dynamic dictionaries induction, even for languages that do not have direct parallel texts, by applying triangulation and using other languages as a bridge. The translations graph contains over 500 k translation pairs in 45 languages. Further, UGARIT allows users to inspect how other users aligned a specific token using the translation pairs search function, which provides detailed results about the different alignment possibilities with a comprehensive visualization. Trilingual alignment, a unique feature of UGARIT, can be used in a variety of ways, such as comparing and visualizing competing translations, evaluating indirect translation by adding the mediating text or even supplementing another translation to a less common language pair. Additionally, UGARIT has been used as a pedagogical tool for vocabulary learning, morpho-syntactic comparisons and learning assessment both in the classroom and in self-study [33]. Finally, UGARIT is designed as a public web tool that does not require any installation, technical expertise, or hardware prerequisites.

In the following sections, we describe the development process and show how manual alignment can be performed in UGARIT. Next, we describe the different visualization approaches used to visualize the translation alignments, the dynamic lexicon search results, and the translation clouds. Finally, we discuss the limitations, possible improvements, and new features we intend to integrate into the next release of UGARIT.

2. Development Process

The development of annotation tools is a challenging task. It requires a deep understanding of the underlying task and the user's needs and an experience in human–computer interaction methods and approaches. Before working on UGARIT, we studied the related tools and defined their limitations, we also consulted numerous research papers and surveys [34–38] that reviewed and analyzed the existing annotation tools and defined design principles and usability recommendations, which helped us to build a primary vision of the tool. The development of UGARIT has been achieved through the close collaboration of researchers from Computer Science, Digital Humanities, Classical Philology, and Translation Studies, aiming to gain a better understanding of users' needs. The development started in 2017 at the Alexander von Humboldt-Chair for Digital Humanities at Leipzig University. We followed the user-centred design principles during the iterative development process (Figure 1) and the usability recommendations for annotation tools [37], leading to a user-friendly, intuitive, easy-to-use tool.

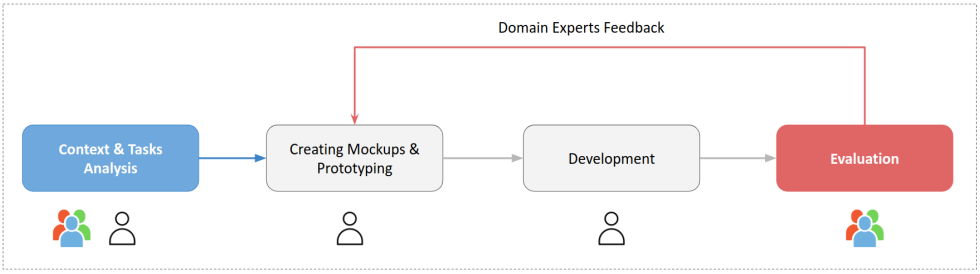

Figure 1. User-Centered Design Process.

The starting point was to define the usage context and analyze the user tasks and requirements; this has been done through interviews and discussions with domain experts who explained their needs and defined the different usage scenarios. Domain experts defined a list of main features that must be provided in the tool: (1) Users can annotate texts in any language, alphabet, or writing system; (2) Users can annotate long text passages; (3) Users can create different alignment types, one-to-one, one-to-many, many-to-one, and many-to-many; (4) Time and clicks required to create the alignment should be minimized to the extent possible; (5) Users can analyze and review their alignments via interactive visualization; (6) Users can download and share their alignments in different formats.

Next, the development team prepared a set of mockups and prototypes that were discussed and approved with the domain experts. Since most of our audience has no technical expertise, we decided to create a public web-based platform-independent tool that is easily accessible and does not require any prior technical knowledge for installation and use.

Later, the first implementation cycle started and resulted in the first product, which has been tested and evaluated by domain experts. The improvement suggestions, problems, and difficulties faced by the domain experts were reported to the development team, who worked to solve these issues and enhance the tool's usability. After multiple development iterations, we could reach a stable and reliable version of that tool.

UGARIT is implemented in Php on the server-side and Javascript on the client-side; MySql and Neo4J are used to store the data. UGARIT supports UTF-8 text format, which allows users to align texts in many languages and different alphabets, such as Aramic, Ancient Greek, and Coptic. It also shows Right-to-Left texts in the correct direction and provides automatic transliteration for non-Latin alphabets languages. The tool is not designed to be used on mobile devices; however, users can still create alignment for short passages with their mobile devices. For best use, it is recommended to use a computer.

Figure 2 summarizes the features and functionalities provided by UGARIT, users can align bi-and trilingual texts; combine their aligned passages in one big text or split a long aligned text into smaller units. Moreover, users can download their alignments, share them, and integrate them into their websites and blogs. Further, UGARIT provides a reading environment for aligned parallel texts, where users can read, search, and compare the aligned passages.

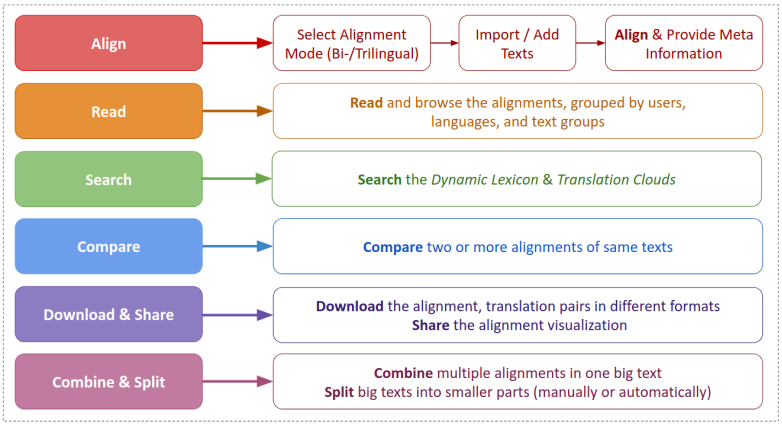

Figure 2. Overview of UGARIT's features and functionalities.

3. Alignment Workflow

Annotation is the process of adding information to a text at some level [39], and we call it *Manual Annotation* when humans perform it. Manual Annotation is a nontrivial task, and it requires intense focus and attention. Like any chore, it becomes boring after a short while. Manual translation alignment, in particular, is a daunting task; it requires moving eyes between two texts in different languages and finding translation correspondences and linking them.

When we designed the UGARIT Interface, we tried to make it easy to use, even for users with no experience with annotation tools. The first step is to create an account on UGARIT or log in if the user has already created an account and verified it. Users have to choose between bi- or trilingual text alignments to start a new alignment. Then, users can upload their parallel texts in plain text format or use the canonical text service (CTS) URNs to import texts from the *Perseus Digital Library CTS* (cts.perseids.org, accessed on 20 October 2021) repository [40], and the languages of the texts must be selected. Next, texts will be tokenized and prepared for alignment.

The alignment process is designed to be as simple as possible; we tried to minimize the mouse clicks needed to create and save the aligned units. In the case of bilingual text alignment, the panel is split into three main columns, two columns for the parallel texts and the right-side column for displaying the aligned translation pairs. A progress bar is located on the top of each language panel to show the alignment coverage so far. UGARIT also offers trilingual alignment, where users can align three parallel texts in three different languages (Figure 3). To align a translation pair, users must select a word/phrase from the original language and then select the corresponding word/phrase in the translated text. To select a token, the user needs to click the token, and then it will be highlighted with green. Clicking a selected token will deselect it and remove the highlighting. The paired tokens will be automatically saved when the user starts to align a new pair or clicks on the *save* icon. On the right side of the editor, users can see all aligned pairs and have the option to edit or delete any pair. The UGARIT editor allows users to create all types of alignment word-to-word (1-1), word-to-phrase (1-N), phrase-to-word (N-1), and phrase-to-phrase (N-N) alignments. The translation pairs list can be exported in XML or tabular format. The resulting translation pairs are automatically stored in the database and then can be exported in XML or tabular format. Furthermore, users can also decide whether the alignment can be publicly visible on the website or keep it private. Finally, users have to provide some information about the texts such as title, translator, and a short description. Once the user saves the alignment, it will appear on the home page in the *New Alignments* panel.

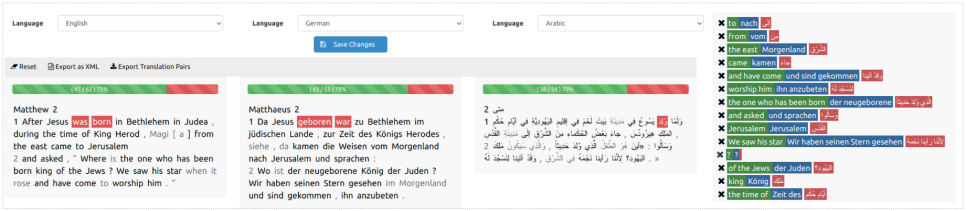

Figure 3. UGARIT trilingual alignment editor shows three side-by-side parallel texts in English, German, and Arabic. The manually created translation equivalents are located on the right.

Figure 4 shows the different alignment types of the translation alignments of an ancient Greek text with its translations in Italian, German, English, and French (The alignments of Homer, Iliad, 21.1-53 Ancient Greek source text and its different translations, alignments are created by Chiara Palladino). The distribution of the alignment types differs according to various factors, including text languages, text type, text genre, translation type, annotator guidelines.

Alignment Guidelines

Annotation guidelines are a key component of any collaborative annotation work; they ensure annotation consistency and reduce annotation errors. The same applies to translation alignment. However, defining alignment guidelines is not a trivial task; it requires a deep understanding of the linguistic structure of both languages and the relations between them. In general, UGARIT does not impose any alignment guidelines, it gives users the freedom to tailor their own guidelines according to their projects or alignment purpose. However, we share some recommendations and refer to common guidelines used for the creation of some alignment gold standards [17,41,42].

We conducted experiments on guidelines for Ancient Greek–English and Persian–English text to measure the impact of the alignment guidelines on the alignment process. The results showed that defining and following alignment guidelines would increase the inter-annotators agreement by at least 10%. However, it is not possible to create guidelines for all language pairs in UGARIT, since an alignment guideline needs to consider multiple aspects of both the target and the source language at once, and should be tailored individually for each language pair by experts.

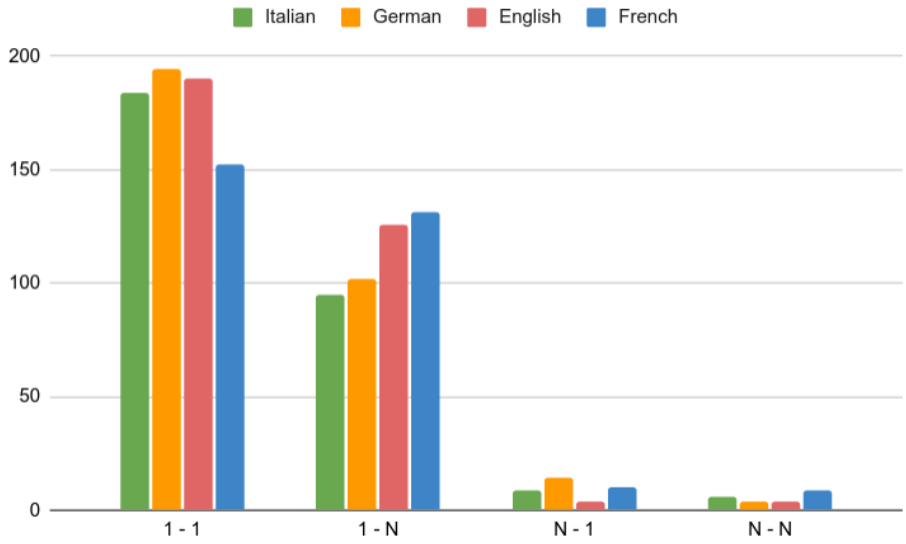

Figure 4. Alignment types of an aligned ancient Greek text with its translations in different languages.

4. Visualization Techniques

Visualization of translation alignment plays an important role in understanding and interpreting the relation between texts and their translations [30]. UGARIT offers various approaches to visualize aligned texts and the derived *Dynamic Lexicon*.

4.1. Languages Graph

The graph is placed on the tool's home page (http://ugarit.ialigner.com, accessed on 20 October 2021) to give users a quick overview of the languages currently hosted and how they are related to each other, as shown in Figure 5.

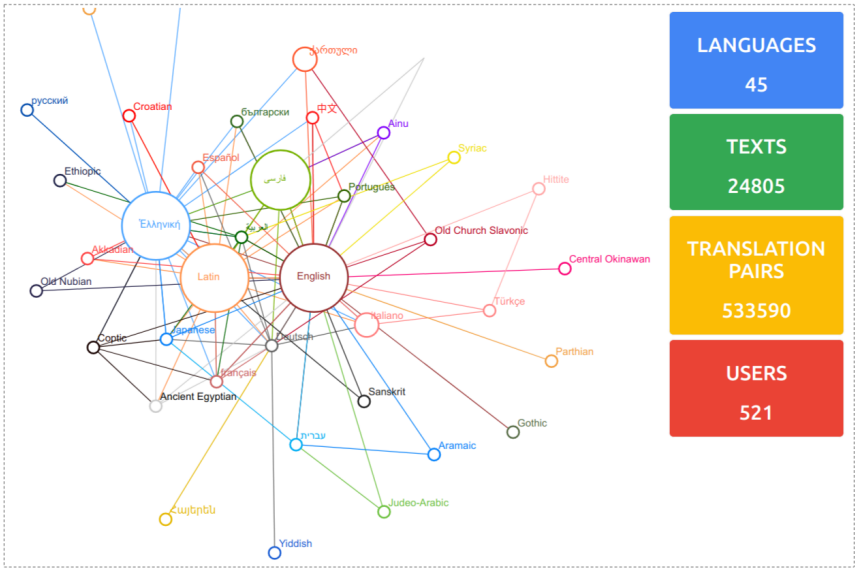

Figure 5. Languages graph.

Each vertex represents a language. The vertex's size reflects the number of texts in this language in the database. Each language is assigned a different color, and vertices are labeled with the language names in the original form. The connection between two vertices means that there are translation alignments between texts in these two languages, and the thickness of the line reflects the number of aligned translation pairs. In Figure 5, clicking

a vertex will show all texts in the language it represents, whereas clicking a link between two vertices will load all aligned texts in these two languages.

4.2. Aligned Texts

The side-by-side view is the most intuitive and straightforward approach to display parallel text alignment at different levels. It is also widely used to visualize collation and mono-language alignments.

Texts are placed alongside each other, as shown in Figure 6. A coloring schema is used to distinguish between aligned and unaligned tokens. Since most tokens are aligned, we used black for aligned tokens and red for the unaligned to draw user attention. The alignment between corresponding tokens on the parallel sides is visualized via highlighting. When the user hovers an aligned word, the hovered word/phrase and the paired tokens will be simultaneously highlighted with red color.

A progress bar is located under each text to give the user an overview of how many tokens are aligned and their percentage; the aligned part is colored with green, whereas the unaligned part is colored with red, as we can see in Figure 6.

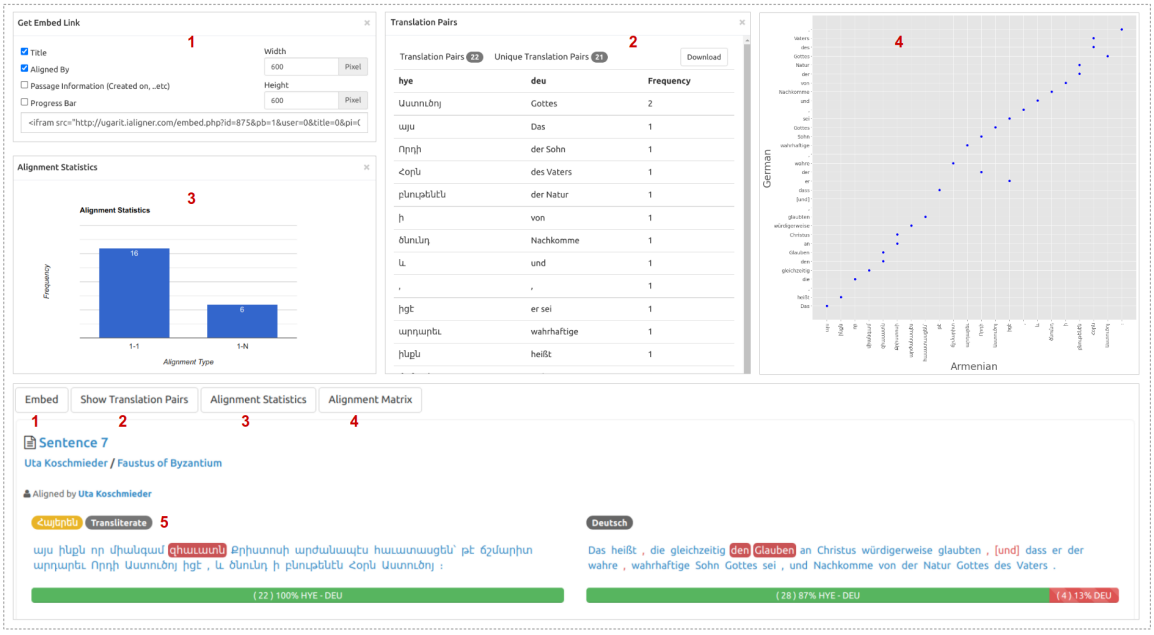

Figure 6. Side-by-side visualization of bilingual aligned texts.

- ***Embed***: This option allows users to generate a link that can be used to share and embed the visualization of their alignments on any blog or website. Furthermore, and users can select which components can be included in the embedded text (progress bar, title, annotator info, text info) (Figure 6(1)).
- ***Translation Pairs***: This option shows an aggregated list of the translation pairs extracted from the aligned text, users can download them in JSON format (Figure 6(2)).
- ***Alignment Statistics***: This option shows statistics of the different alignment categories. This chart provides valuable information on the alignment quality, user's language knowledge, and the relation between the aligned languages (Figure 6(3)).
- ***Alignment Matrix***: with this option, users can view the alignment in the form of a grid: the source text tokens are located on the horizontal axis, and the translation tokens are located on the vertical axis. The blue dots represent alignment between the corresponding column and row tokens. The diagonal dots indicate one-to-one alignments, whereas the vertical ones indicate the one-to-many (Figure 6(4)).
- ***Transliteration***: UGARIT contains texts in various languages with different alphabets. For better readability, especially for new language learners, UGARIT offers an auto-

matic transliteration for non-Latin alphabets languages, which is visible when the pointer hovers the aligned word. This feature is currently available for Greek, Arabic, Persian, Armenian, and Georgian (Figure 6(5)).

- ***Combine/Split Aligned Texts***: UGARIT enables users to merge multiple aligned texts into one bigger text. It also lets users split a long aligned text into smaller units (sentences/paragraphs). This feature is beneficial since annotators prefer to align long paragraphs over short ones to avoid copying and pasting them multiple times. On the other hand, splitting long aligned text is useful when users want to create shorter aligned units that can be used later to train or evaluate machine translation or translation alignments models.

4.3. Translation Pairs & the Dynamic Lexicon

UGARIT provides a search function to enable users to look for a word or phrase and get detailed information about how this word/phrase occurs in the UGARIT texts collection and how it is translated in different languages. Moreover, UGARIT visualizes the dynamic lexicon search results in two approaches:

Tree View is a classic branching view that enables users to navigate the hierarchy to filter the results set until they reach the desired subset. Figure 7a shows how UGARIT visualizes the search results as three levels tree view. The query word is located at the first level as the root word, corresponding translations are grouped by languages, and languages are placed as nodes at the second level. The aligned translations are placed at the third level. The language nodes are initially collapsed for clarity, and users can expand them to explore the translations by clicking the language label.

Radial Cluster Dendrogram is similar to the tree view mentioned above; it shows the translation equivalents in the form of a rooted tree and all leaf nodes are placed at the same depth (Figure 7).

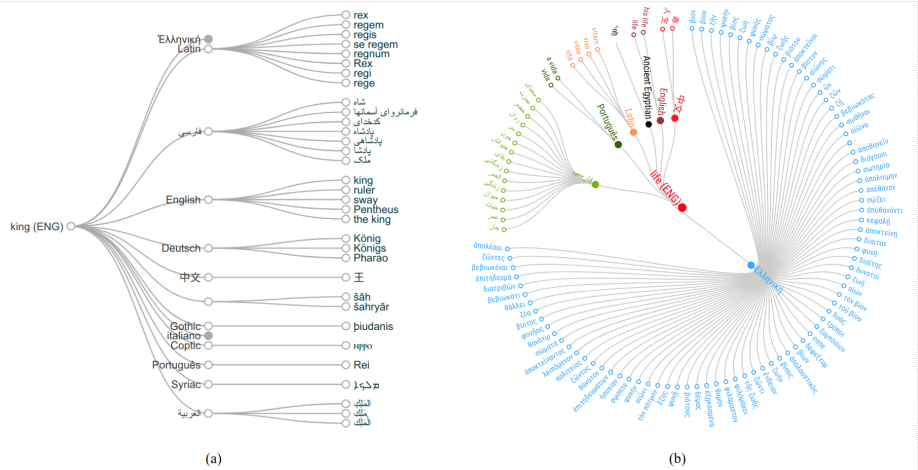

Figure 7. Visualization of TPs Search Results. (**a**) Tree View, (**b**) Radial Cluster Dendrogram View.

These two views illustrate the idea of the *Dynamic Lexicon* [7], which is an automatically derived lexicon from parallel aligned texts at the word/phrase level. *Dynamic Lexicon* uses triangulation to create new dynamic lexica using pivot languages based on the assumption that two words/phrases are likely to be translations if they are translations of the same word in a third language [43]. In the two examples shown in Figure 7, triangulation considers all leaf nodes synonyms (if they are in the same language) or translation equivalents since they share the exact English translation.

5. Translations Graph

UGARIT is characterized by its simplicity and ease of use, and for this reason, it has been used by many users and in several projects, which led to continuous growth in the

number of aligned texts and translation pairs. Though the traditional relational database architecture implemented in UGARIT is no longer satisfactory, we realized that we could achieve more benefits if we modelled our data more sophisticatedly.

After experimenting with different models, we came to the conclusion that the graph data structure fits our data best. The proposed model can be described as follows:

Let $G = (V, E)$ be a translations-graph. where V is a set of all words/phrases in the the UGARIT aligned texts, and E is a set of all translation pairs. Let $v_1 \in V$ be a word/phrase in language l_1, and v_2 be a word/phrase in language l_2. $\{v_1, v_2\}$ be a translation pair that occurs n times in the dataset.

Each node $v \in V$ has four attributes, an internal id id, label t, language l, and n the frequency of this word/phrase in the whole dataset $v\{id, t, l, n\}$.

The *translation* relation between $v_1\{id_1, t_1, l_1, n_1\}$ and $v_2\{id_2, t_2, l_2, n_2\}$ can be represented as an edge $e \in E$, e encapsulates a single value n which indicates the frequency of this translation pair in the dataset. We have used the Neo4j database [44] to model the translations graph.

Translation Clouds

Visualizing the translation graph as a connected graph was not possible due to the large number of nodes and crossing edges, which would affect the translations' clarity and readability. The solution was to keep the labels and remove the lines that connect them, resulting in a word cloud representation.

Translation cloud shows the translation equivalents to a given word, these equivalents are extracted from the translation graph up to the fifth level. Figure 8 shows the translations cloud of the root word *city*.

Figure 8. Translations cloud of the root word *city*.

The view consists of three components, the translation cloud, the languages bar above it, and the concrete translation path on the right side. The search word is located at the cloud center, surrounded by its equivalents in other languages. Further, translation

equivalents are grouped by language, and every language is assigned a unique color. However, languages are listed in the languages bar; clicking on a language label will hide all translation equivalents in other languages and keep the ones in the selected language. Labels are displayed in different font sizes according to the distance to the root word. Since the translations cloud visualizes connected translation equivalents up to the fifth level, the distance varies between 1 to 5; the lower the distance value, the higher the font size. Moreover, clicking a translation label will show the path between the selected word and the root. Additionally, a tooltip is used to show information about the label, such as language and frequency.

Using the Neo4j graph database at the back-end to store the translation pairs has shown many advantages; it reduced the response time of search queries, enabled us to perform complex queries, and facilitated the dynamic lexicon production.

6. UGARIT in Research and Pedagogy

The complexities involved in the operation of translation alignment are often a reflection of the dynamic relationship between original and translation, and between different languages. UGARIT provides an environment where the user can engage with those complexities in a very analytical way, on both small and big scale.

On a small scale, UGARIT offers an immersive and visually powerful environment, where the reader can analytically compare texts token by token, and at the same time observe the results through an interactive visualization. The comparison of parallel texts becomes a systematic operation, which encourages reflection regarding the interplay between two languages, the meaning of specific words, and overall the (im)perfect matching of words and expressions. This is also an exercise in cultural dialogue and reflection, not only upon the language(s) but upon the civilization that used it to reflect its values. Moreover, the opportunity to publish the results online provides a way to be part of a broader conversation on the reception and significance of a text over time. For these reasons, UGARIT is currently used in studies on translation and reception and language teaching, particularly for Ancient Greek and Latin, across the world. In pedagogy, translation alignment is often integrated with grammatical and syntactical observations to emphasize the complex interplay between the language of the translation and the target language, and students are assigned various alignment tasks and exercises to empower the analytical approach to the text [33,45]. On a bigger scale, UGARIT also provides manually aligned parallel corpora across languages that had never been compared before, such as Coptic, Ancient Greek, Arabic, Persian, Latin, Egyptian, Georgian, etc. The analysis of these aligned datasets and recurrent patterns in word matching can provide insights into how cross-linguistic and cross-cultural dynamics are affected by different language structures, cultural differences, text genres, and even language proficiency [46].

Since the tool was made public, the number and variety of languages included by the users has steadily increased and has gone far beyond the original intent: at the moment this paper is being written, 36 languages are included in UGARIT, and there are 295 active users, and about 23,500 parallel texts.

UGARIT has been used as a pedagogical tool for vocabulary learning, morpho-syntactic comparisons and learning assessment both in the classroom and in self-study. One case in point is an online course for teaching Ancient Greek in Persian speakers [47], where UGARIT was used extensively. After a five-minute tutorial, all participants of the course were able to register and use UGARIT for aligning translation successfully. During this course, UGARIT was used as a reading environment for parallel texts by the educator during the class, as a vocabulary learning tool by the students, and as a assessment/practice environment where the educator could assess and examine the students' understanding of the text. Furthermore, the trilingual alignment provided a suitable environment for using a third language as a bridge between the source language and the target language, which in case of the Ancient Greek course would be English.

In addition to the pedagogical aspect, the trilingual alignment is a practical tool for research purposes, providing valuable information on complex texts, particularly through terminology extraction. In Figure 9, an obscure term from the inscription of Shapur I at Ka'ba-i Zardušt (ŠKZ) is shown along with its English and Greek equivalents throughout the text, presenting both morphological and semantic variations in the Greek version of the inscription.

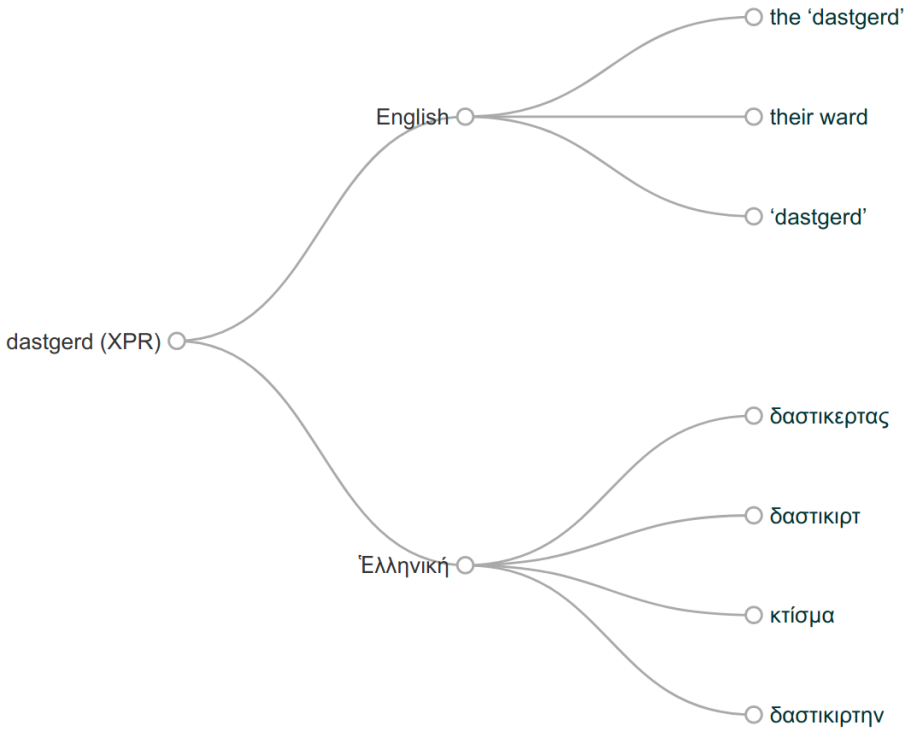

Figure 9. Visualization of translation equivalents of the *Parthian* word *dastgerd*.

7. Future Work

When we created UGARIT, we aimed to visualize the automatically aligned texts and collect training data for automatic alignment systems. Still, UGARIT has also been used in many research projects [48,49] and for different purposes, especially teaching and learning historical languages [50]. Experts, who have been using UGARIT regularly, have provided us with some ideas and improvement suggestions to enhance the usability of UGARIT. Therefore, new features and functionality should be implemented and developed to keep pace with users' needs. We can sum them up as follows:

- *User roles:* the next version of UGARIT will offer different user roles such as *expert, instructor, student*, which would help create accurate training data by considering the alignments created by experts and instructors, since they are supposed to produce correct and precise alignments. In contrast, students in the learning phase could make some alignment mistakes, and these mistakes should not affect the accuracy of the dynamic lexicon and training datasets;
- *Teaching:* Further, experts and instructors will be able to create groups, add students to the groups, and create assignments. Instructors can upload these assignments in the form of plain parallel texts; students will be asked to align them with deadlines, with the possibility of uploading the correct alignment to allow the system to evaluate the assignments automatically and give notes to every student;
- *Alignments sharing and exporting:* in the current version of UGARIT, users can export their results in XML format only. The next version will offer other formats such as

JSON and CSV to facilitate the reuse of the alignment in other applications or for other purposes;
- *Automatic alignment:* we are currently developing an automatic alignment system and planning to offer an automatic alignment option for texts in specific languages, such as Ancient Greek–English and Latin–English, or at least supporting the users with alignment suggestions to reduce the time required to align long texts;
- *Collaborative alignment:* in the current version, users can only align their texts; however, the next version will provide an option for the collaborative alignment of long texts where multiple users can work on the same text.

Author Contributions: Conceptualization, T.Y., C.P. and M.F.; Software, T.Y.; Visualization, T.Y.; Writing—original draft, T.Y.; Data curation & Investigation; T.Y., C.P., F.S. and M.F.; Writing—review & editing, T.Y., C.P., F.S. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: The publication of this paper is supported by the Open Access Publishing Fund of Leipzig University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kay, M.; Röscheisen, M. Text-Translation Alignment. *Comput. Linguist.* **1993**, *19*, 121–142.
2. DeNero, J.; Klein, D. Tailoring word alignments to syntactic machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 17–24.
3. David, Y.; Grace, N.; Richard, W. Inducing multilingual text analysis tools via robust projection across aligned corpora. In Proceedings of the First International Conference on Human Language Technology Research, San Diego, CA, 18–21 March 2001; pp. 1–8.
4. Padó, S.; Lapata, M. Cross-lingual annotation projection for semantic roles. *J. Artif. Intell. Res.* **2009**, *36*, 307–340. [CrossRef]
5. Durrani, N.; Koehn, P. Improving machine translation via triangulation and transliteration. In Proceedings of the 17th Annual Conference of the European Association for Machine Translation, Dubrovnik, Croatia, 16–18 June 2014; pp. 71–78.
6. Wu, D.; Xia, X. Learning an English-Chinese lexicon from a parallel corpus. In Proceedings of the First Conference of the Association for Machine Translation in the Americas, Cuernavaca, Mexico, 10–14 October 1994.
7. Yousef, T.; Berti, M. The digital fragmenta historicorum graecorum and the ancient greek-latin dynamic lexicon. In *Corpus-Based Research in the Humanities (CRH)*; Institute of Computer Science: Warsaw, Poland, 2015; p. 117.
8. Brown, P.F.; Cocke, J.; Della-Pietra, S.A.; Della-Pietra, V.J.; Jelinek, F.; Lafferty, J.D.; Mercer, R.L.; Rossin, P. A statistical approach to machine translation. *Comput. Linguist.* **1990**, *16*, 76–85.
9. Gale, W.A.; Church, K.W. A program for aligning sentences in bilingual corpora. In Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL), Berkeley, CA, USA, 18–21 June 1991.
10. Moore, R.C. Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, 6–12 October 2002, Proceedings*; Richardson, S.D., Ed.; Springer: Berlin, Heidelberg, 2002, Volume 2499.
11. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
12. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv* **2019**, arXiv:1911.02116.
13. Jalili Sabet, M.; Dufter, P.; Yvon, F.; Schütze, H. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Online, 2020; pp. 1627–1643. [CrossRef]
14. Dou, Z.Y.; Neubig, G. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; Association for Computational Linguistics: Online, 19–23 April 2021; pp. 2112–2128.
15. Bojar, O.; Prokopová, M. Czech-English Word Alignment. In Proceedings of the LREC, Genoa, Italy, 22–28 May 2006; pp. 1236–1239.

16. Graça, J.; Pardal, J.P.; Coheur, L.; Caseiro, D. Building a Golden Collection of Parallel Multi-Language Word Alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*; European Language Resources Association (ELRA): Marrakech, Morocco, 28–30 May 2008.
17. Mareček, D. Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus. Master's Thesis, Charles University, Prague, Czech Republic, 2008.
18. De Pauw, G.; Wagacha, P.W.; de Schryver, G.M. The SAWA corpus: A parallel corpus English-Swahili. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on 'Language Technologies for African Languages'*; Association for Computational Linguistics: Athens, Greece, 30 March 2009; pp. 9–16.
19. Holmqvist, M.; Ahrenberg, L. A Gold Standard for English-Swedish Word Alignment. In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), Riga, Latvia, 11–13 May 2011; Northern European Association for Language Technology (NEALT): Riga, Latvia, 2011; pp. 106–113.
20. Melamed, I.D. Manual Annotation of Translational Equivalence: The Blinker Project. *arXiv* **1998**, arXiv:cmp-lg/9805005.
21. Grimes, S.; Li, X.; Bies, A.; Kulick, S.; Ma, X.; Strassel, S. *Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC*; European Language Resources Association (ELRA): Valletta, Malta, 2010.
22. Benner, D. A Tool for a High-Carat Gold-Standard Word Alignment. In Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), Gothenburg, Sweden, 26 April 2014; pp. 80–85.
23. Caseli, H.M.; Feltrim, V.D.; Nunes, M.G.V. *TagAlign: Uma Ferramenta de Pré-Processamento de Textos*; Série de Relatórios do NILC. NILC-TR-02-09 Junho; NILC: Berkeley, CA, USA, 2002.
24. Germann, U. Yawat: Yet another word alignment tool. In Proceedings of the ACL-08: HLT Demo Session, Columbus, OH, USA, 10 January 2008; pp. 20–23.
25. Almas, B.; Beaulieu, M.C. Developing a New Integrated Editing Platform for Source Documents in Classics. *Lit. Linguist. Comput.* **2013**, *28*, 493–503. [CrossRef]
26. Gilmanov, T.; Scrivner, O.; Kübler, S. SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer. In Proceedings of the LREC, Reykjavik, Iceland, 26–31 May 2014; pp. 2913–2919.
27. Barreiro, A.; Raposo, F.; Luís, T. CLUE-Aligner: An alignment tool to annotate pairs of paraphrastic and translation units. In Proceedings of the 10th Language Resources and Evaluation Conference (LREC), Portorož, Slovenia, 23–28 May 2016; pp. 7–13.
28. Almas, B.; Berti, M. Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors. In Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities, DH-CASE '13, Florence, Italy, 10 September 2013; Association for Computing Machinery: Florence, Italy, 2013. [CrossRef]
29. Callison-Burch, C.; Talbot, D.; Osborne, M. Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21–26 July 2014; pp. 175–182. [CrossRef]
30. Yousef, T.; Janicke, S. A Survey of Text Alignment Visualization. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 1149–1159. [CrossRef] [PubMed]
31. Jänicke, S.; Franzini, G.; Cheema, M.F.; Scheuermann, G. Visual text analysis in digital humanities. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2017; Volume 36, pp. 226–250.
32. Bamman, D.; Crane, G. Measuring Historical Word Sense Variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*; Association for Computing Machinery: New York, NY, USA, 2011; pp. 1–10. [CrossRef]
33. Palladino, C.; Foradi, M.; Yousef, T. Translation Alignment for Historical Language Learning: A Case Study. *Digit. Humanit. Q.* **2021**, *15* . Available online: https://www.proquest.com/openview/e048d32e8e991c67282c3fbda5c1f0d4/1?pq-origsite=gscholar=5124193 (accessed on 4 November 2021).
34. Neves, M.; Leser, U. A survey on annotation tools for the biomedical literature. *Brief. Bioinform.* **2014**, *15*, 327–340. [CrossRef] [PubMed]
35. Ide, N. Introduction: The Handbook of Linguistic Annotation. In *Handbook of Linguistic Annotation*; Ide, N., Pustejovsky, J., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp. 1–18. [CrossRef]
36. Finlayson, M.A.; Erjavec, T. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 167–191.
37. Burghardt, M. Usability Recommendations for Annotation Tools. In *Proceedings of the Sixth Linguistic Annotation Workshop*; Association for Computational Linguistics: Jeju, Korea, 12–13 July, 2012; pp. 104–112.
38. Burghardt, M. Engineering Annotation Usability—Toward Usability Patterns for Linguistic Annotation Tools. Ph.D. Thesis, University of Regensburg, Regensburg, Germany 2014.
39. Petrillo, M.; Baycroft, J. *Introduction to Manual Annotation*; Fairview Research: Haven, CT, USA, 2010.
40. Babeu, A. *The Perseus Catalog: Of FRBR, Finding Aids, Linked Data, and Open Greek and Latin*; Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution; Berti, M., Ed.; De Gruyter Saur: Berlin, Germany; Boston, MA, USA, 2019; pp. 53–72. [CrossRef]

41. Li, J.J.; Kim, D.I.; Lee, J.H. Annotation Guidelines for Chinese-Korean Word Alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*; European Language Resources Association (ELRA): Marrakech, Morocco, 26 May 2008.
42. Kholidy, H.A.; Chatterjee, N. Towards developing an Arabic word alignment annotation tool with some Arabic alignment guidelines. In Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, 29 November–1 December 2010; pp. 778–783. [CrossRef]
43. Ács, J. Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*; European Language Resources Association (ELRA): Reykjavik, Iceland, 26–31 May 2014; pp. 1938–1942.
44. Database, N.G. Graph Modeling Guidelines. 2019. Available online: http://neo4j.com/developer/guide-data-modeling (accessed on 2 November 2019).
45. Palladino, C. Reading Texts in Digital Environments: Applications of Translation Alignment for Classical Language Learning. *J. Interact. Technol. Pedagog.* **2020**, *18*, 724–731.
46. Palladino, C.; Yousef, T. We Want to Learn All Languages! Digital Classicist Seminar London. 2021. Available online: https://www.youtube.com/watch?v=R2Ms6yAMZss (accessed on 27 October 2021).
47. Shamsian, F. Digital Classics and Learning Greek in Iran, Sunoikisis Digital Classics. Thursday, 14 May 2020. Available online: https://www.youtube.com/watch?v=ernL2sRGJ-U (accessed on 27 October 2021).
48. The Digital Rosetta Stone Project. 2019. Available online: https://rosetta-stone.dh.uni-leipzig.de/ (accessed on 2 November 2019).
49. Shukhoshvili, M. Methodology of Translation Alignment of Georgian Text of Plato's "Theaetetus". *Int. J. Lang. Linguist.* **2017**, *4*, 63–69.
50. Foradi, M.; Palladino, C.; Shamsian, F. Confronting Complexity of Babel in a Global and Digital Age. In *DH2019: Digital Humanities Conference, Book of Abstracts*; Utrecht University, Utrecht, The Netherlands, 2019; pp. 127–138.