Ugarit: Translation Alignment Visualization

Tariq Yousef

NLP Group, Institute for Computer Science, University Of Leipzig, Germany
E-mail: tariq@informatik.uni-leipzig.de

***Abstract*—In this paper we introduce Ugarit a web-based tool for manual translation alignment of parallel texts, the aim was to build a user interface to create training data in form of translation pairs to be used later for an automatic translation alignment system at word/phrase level, the tool now is widely used as reading environment for parallel texts and a languages learning tool. The paper gives a short overview of the visualization techniques used to present the alignment results and shows how the translation graph derived from the aligned translation pairs.**

***Index Terms*—Translation Alignment, Data Visualization, Manual Alignment, Translations Graph.**

I. Introduction

Translation alignment is one of the most popular applications of Digital Humanities and Natural Language Processing. It is defined as the comparison of two or more texts in different languages called parallel texts to find which units of the source and target texts correspond together [1]. we can distinguish between various levels of translation alignment according to the text unit to be aligned, it can be document, paragraph, sentence, or word/phrase alignment. Bilingual text alignment is an essential task in statistical Machine translation, several automatic approaches has bees developed to perform the alignment at different levels such as [16] for sentence level alignment and IBM models for word alignment. Most of them use unsupervised statistical methods to create alignment probabilities distribution between the source and target text units.

The accuracy of the automatic methods varies according to various factors such as text type, text length, size of the corpus, and translation accuracy. In contrast, Manual alignment produces more accurate results since it is performed by scholars and experts, but it is expensive in term of time and resources. The lack of user-friendly annotation tools for translation alignment and the need for accurate training data to perform the automatic alignment drove us to create Ugarit. Originally, Ugarit was developed to collect training data for the implementation of statistical machine translation system for historical languages, mainly Ancient Greek, Latin, and Persian, for which few to none aligned data sets exist. Ideally, historical languages are closed systems with a finite number of words and very limited change in the foreseeable future. Therefore, it should be possible to create adequately efficient automated methods of statistical machine alignment based on a relatively small training data set.

The development of Ugarit started in 2017 at the AvH Chair for digital humanities at university of Leipzig, under supervision of Prof. Gregory Crane and collaboration with Chiara Palladino and Maryam Foradi.

Ugarit is crowd-sourcing project enables users to create translation alignments at word level, and the resulting translation pairs can be re-used in future machine or human translations or to create dynamic lexica and translation memory.

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.

In next section we review similar and related works, then we explain how the manual alignment can be done in Ugarit, next we show the various visualization techniques used by Ugarit to visualize the translations alignments and the dynamic lexicon search results, then we present the translations graph. Finally, we discuss the possible improvements and new features we intend to add to the next release of Ugarit.

II. Related Works

The Blinker Project [17] has developed the first annotation tool for manual text alignment to align different versions of the Bible in French and English at word level. Alpheios [4] can be considered as the first public web-based manual alignment tool for translation alignment. Alpheios [3] offers two options for the visualization of aligned texts, the side-by-side approach and the interlinear text approach.

However, there are several available softwares for automatic text alignment at sentence level such as **YouAlign**[1], and Moore's bilingual sentence aligner [9]. Giza++ [5] uses IBM models to create automatic alignments at word level.

Visualization of aligned texts was the subject of interest and research in the recent years, many tools have been developed for this purpose. Some tools such as JuXtacommons [13] and Versioning Machine use side-by-side approach to visualize the alignment between texts in the same language at word level. Other tools like TRAViz [11] and Stemmaweb [10] use variant graph to represent the alignment between text variants.

[1] www.youalign.com

Fig. 1: Ugarit Texts Alignment Editor

Alignment Map [20] uses the parallel view to visualize translation alignment, it shows the base text and its translation, the blue connecting lines indicate alignment, the thickness of the line reflects the length of the aligned segments in words. Iteal [12] introduced the Meso reading to visualize the alignment between parallel texts at line and word level. Other tools like CATView [6] and iAligner [14] use the tabular view approach to show the positions of similarity and difference between parallel texts.

III. Workflow

When we designed Ugarit Interface we tried to make it easy to use even for users who has no experience with annotation tools. Users need to create account on Ugarit and sign in to be able to start a new alignment. Users have the possibility to upload their own corpus in plain text format, or use the canonical text service (CTS) URNs to import texts from the Perseus Digital Library CTS [2]repositories [19]. then the languages must be selected.

Texts will be tokenized and prepared for the next step, 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 tokens. To align a translation pair, users have to select the word/phrase from the original language and then select the corresponding tokens in the translated text (Fig 1). The paired tokens will be saved automatically when user starts to align new pair. On the right side of the editor, users can see all aligned pairs and have the option to edit or delete any pair. Ugarit editor allows user to create all types of alignment one-to-one, one-to-many, many-to-one, and many-to-many alignment. Resulting pairs are automatically stored in the database, and can then be exported in XML or tabular format. Users are asked to provide some information about the texts such as title, translator, description,.. etc, Users can also decide whether the alignment can be publicly visible on the website or not. Once user saves the alignment, it will appear on the home page in the "New Alignments" panel.

IV. Visualization Techniques

Visualization of translation alignment plays a great role to understand and interpret the relation between parallel texts and how the aligned patterns are relevant. Ugarit uses different visualization approaches to give the users an overview of the languages and texts hosted by Ugarit and to visualize aligned texts and translation pairs in the search results.

A. *Languages Graph*

The graph is placed on the home page of the tool [3] to give users a quick overview of the languages currently hosted and how they are related to each other as you can see in Fig 2.

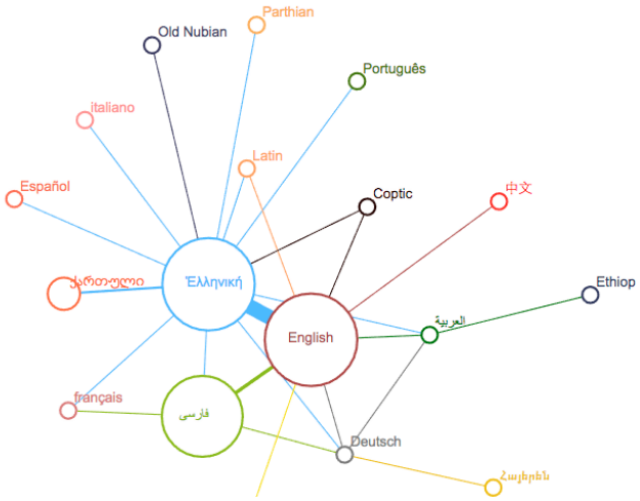

Fig. 2: Languages Graph.

Each vertex represents a language, the size of the vertex reflects the number of texts in this language in the database, each language is assigned a different color, vertices are labelled with the language names in the original form. The connection between two vertices means that there are translations alignment between texts in these two languages, and the thickness of the line reflects the number of aligned translation pairs. In Figure 2 we notice the thick blue line between English and Ancient Greek, because Ugarit contains a huge amount of English-Ancient Greek automatic aligned texts at word level provided be Perseus Digital Library [5].

Clicking on a vertex will load all texts in the language it represents. Clicking on the a line between two vertices will load all aligned texts in these two languages

B. *Aligned Texts*

Side-by-side visualization is the simplest and most used technique to display the alignment of parallel texts at different levels.

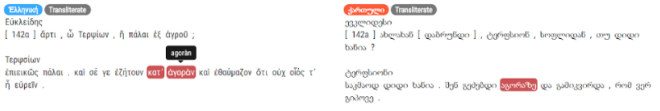

Fig. 3: Two texts side-by-side Visualization

[2]cts.perseids.org

[3]Ugarit.ialigner.com

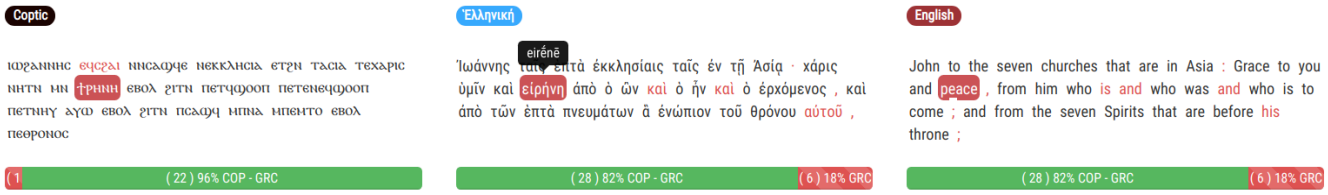

Fig. 4: Three texts side-by-side Visualization

Texts are placed alongside each other as we can see in Fig 3 and Fig 4, a coloring schema is used to distinguish between aligned and unaligned tokens, since the majority of tokens are aligned we used the red color for the unaligned tokens to draw user attention, whereas black color is used for aligned tokens. The alignment between aligned tokens on the parallel sides is visualized via highlighting, when user moves the pointer over an aligned word, the hovered word/phrase and the paired tokens will be simultaneously highlighted with red color.

Since Ugarit hosts translation alignments in various languages with different alphabets, and for better readability, Ugarit offers an automatic transliteration for non-Latin alphabets languages, which is visible when the pointer hovers on the aligned word. This feature is currently available for Greek, Arabic, Persian, Armenian, and Georgian.

A progress bar is located under each text to give the user an overview of how many tokens are aligned and their Their percentage, the aligned part is colored with green, whereas the unaligned part is colored with red as we can see in Fig 4.

C. *Dynamic Lexicon*

The dynamic lexicon [7] is an automatically derived lexicon from parallel aligned texts at word/phrase level. Triangulation is used to create new dynamic lexica using pivot languages based on the assumption that two expressions are likely to be translations if they are translations of the same word in a third language [8].

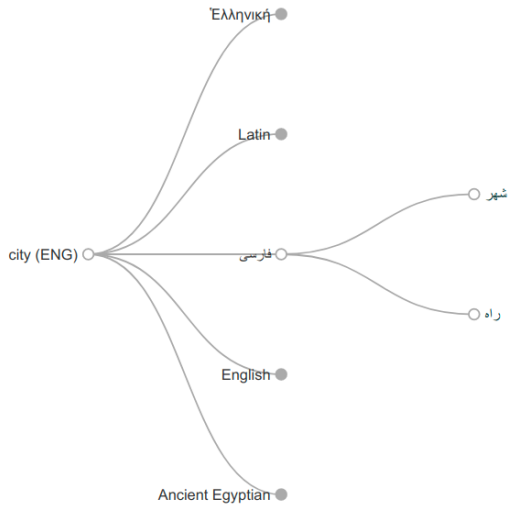

Fig. 5: Tree View Visualization of The Dynamic Lexicon

Ugarit offers a search function to enable users to look for a word or phrase and get a detailed information about how this word/phrase occurs in the corpora and how it is translated in different languages. Ugarit visualizes the dynamic lexicon search results in two approaches:

1) ***Tree View****:* is a classic branching view which enables users to navigate through the hierarchy to filter the results set until reaching the desired subset.

Figure 5 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 languages nodes are initially collapsed for clarity, users can expand the nodes to explore the translations by clicking the languages nodes.

2) ***Radial Cluster Dendrogram****:* is a visualization approach to show hierarchical data in form of rooted tree. Figure 6

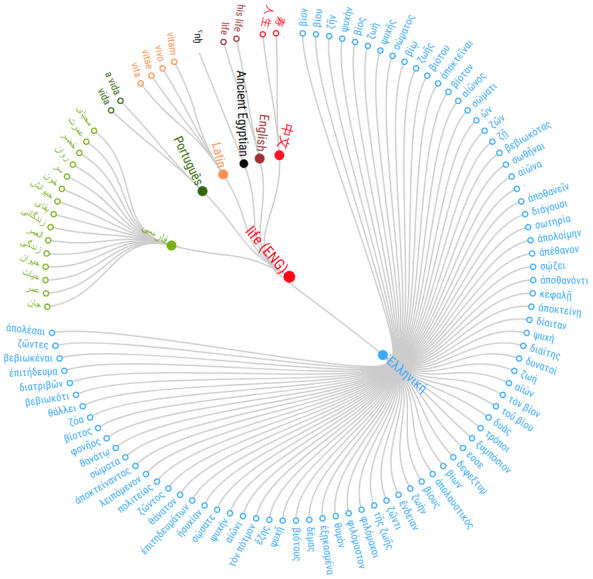

Fig. 6: Radial dendrogram visualization of the dynamic lexicon

show how Ugarit shows the translation of a given word in form of radial dendrogram. The search term is placed in the center of the circle as the root of the tree, the children are the languages are placed in the inner-most ring. The leaf nodes are the translations of the search term clustered by the languages

are place with equal depth. Each language is colored with different color.

V. Translations Graph

Because of the huge amount of translation pairs available on Ugarit, we needed to model the relation between word/phrases in more sophisticated way to achieve the best benefit of it. After experimenting different models, we came to the conclusion that the graph data structure fit our data best. The model can be described as follows:

Let $G = (V, E)$ is 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 $v_1 \in V$ is a word/phrase in language l_1, and v_2 is word/phrase in language l_2. Suppose that the translation pair $\{v_1, v_2\}$ occur n times in the data set.

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

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 an edge $e \in E$, e encapsulates a single value n which indicates the frequency of this translation pair in the data set.

We have used Neo4j database [18] to model the translations graph. Since Neo4j doesn't support indirect relation between nodes, we had to represent each translation pair as two nodes connected with two directed edges as we can see in figure 7.

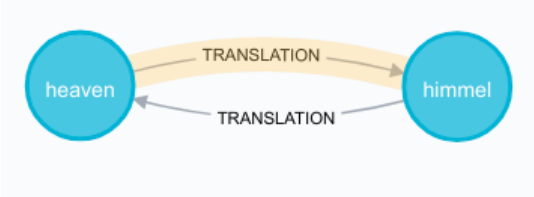

Fig. 7: Graph Representation of a Translation Pair

A. Translation Clouds

visualizing the translations graph as connected graph was not possible because of so many crossing edges which will affect the clarity and readability of the translations. Th solution was to keep the labels and remove the lines between them which produced a word cloud representation. Figure 9 shows the translations cloud of the root word *city*. The search word is located at the center of the cloud surrounded with its equivalents in other languages. Translations are grouped by languages and every language is assigned an unique color. The label's font size varies according to the distance between the root word and the current word. Since the translations cloud visualizes connected translations up to the fifth level, the distance varies between 1 to 5, the lower the distance value, the higher the font size. A tooltip is used to show information about the label such as language and frequency. The path between any translation and the root word will be displayed on the right side when a label is clicked.

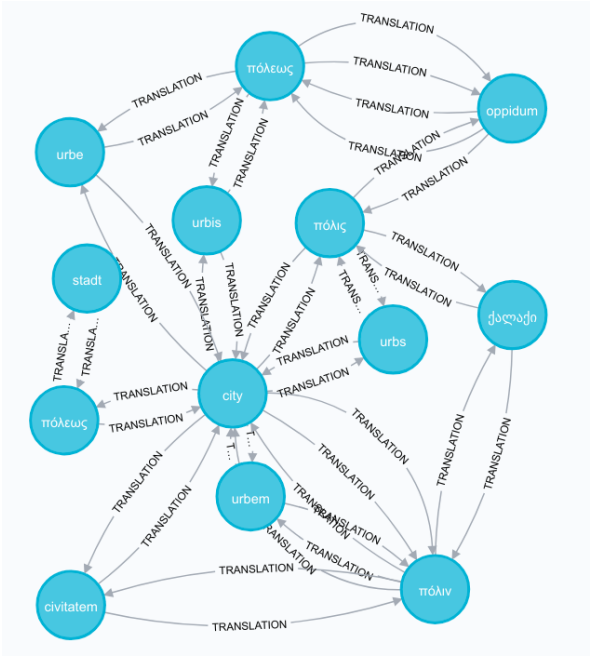

Fig. 8: Translations Graph

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

VI. Future Work

The next version of Ugarit will offer different user roles such as *expert, instructor, student..etc*, that would be very helpful for creating accurate training data, since experts and instructors are supposed to create correct and accurate alignments whereas students in the learning phase could make some alignment mistakes, and these mistakes should not affect the accuracy of the dynamic lexicon and training data sets.

Experts and instructors will be able to create classes, add students to the classes, and create assignments in form of unaligned texts and let students to align them with deadlines, with the possibility of uploading the correct alignment to let the system to evaluate the assignments automatically and give notes to every students. In the current version, users can only align their own texts, the next version will provide an option for collaborative alignment of long texts or working on collaborative text groups.

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, and users will be able to share their alignments and visualization via unique URLs. A clear and rich documentation will be provided in form in form of texts, tutorials and videos to make the use of s easy as possible.

The next version will also offer the automatic alignment option, the system checks if there are enough training data to run the automatic alignment, Users will be able to edit and