A Graph-based Approach for Learner-tailored Teaching of Korean Grammar Constructions

1st Mikhail G. Akimov
Laboratory of Oxylipins
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS
Moscow, Russia
akimovmike@gmail.com

2nd Ekaterina D. Loginova
Faculty of Economics and Business Administration, Department of Marketing
Ghent University
Ghent, Belgium
ekaterina.loginova@ugent.be

3rd Maxim M. Musin
Rebels.media
Belgium
maxim@musin.cc

***Abstract*—Foreign language learning on an intermediate level is often a complicated task, as it requires acquisition not only of vocabulary and language rules but of context-dependent meanings of words. This is especially relevant for Category IV languages like Korean, in which the same tokens could be both words and grammar tags. The textbook adapted versions of words and contexts often fail to capture the existing complexity, while the real world examples may be too hard for a novice and even an intermediate level learner. In addition, the particular learner may be familiar with some functions and contexts for a particular word, but not with the other ones. To alleviate this complexity problem, we propose a semantic graph based personalized tutoring system. The learning corpus is constructed using real-world sentences from a newspaper, which are translated using an automated service and processed with NLP techniques to extract token functions. A graph is used to track word and grammar construct context and thus find similar and dissimilar word use cases, as well as for the estimation of sentence complexity. The system then shows words and grammar constructs from real-world sentences to learners and records their understanding in each context. The collected context dependent understanding data together with the sentence complexity estimation are then used to estimate the learner's level and tailor the sentence set accordingly. The resultant approach could be extended to the tutoring of context-dependent meanings in other languages.**

***Index Terms*—semantic graph, NLP, personalized learning, Korean grammar constructions, context dependent word meaning**

I. Introduction

Computer-assisted language learning is the field of study that deals with bringing an algorithmic approach to second language acquisition. One of the primary aims is to build intelligent tutors which are capable of providing a truly personalized and contextualized learning process.

There is a growing appeal for the use of the data-driven language learning techniques. One of the main reasons is the abundance of language data available in a digital format. Most of the current work is however focused on essay correction and spaced repetition techniques. There are very few approaches which utilize natural language processing (NLP) for illuminating the semantic and syntactic workings of the language.

This turns out to be even more problematic for non-European languages. For instance, in Korean, meanings of the words and their functions are highly context-dependent. The tokens could be both words and grammar tags. The textbook adapted versions of words and contexts often fail to capture the existing complexity, while the real world examples may be too hard for a novice and even an intermediate level learner. In addition, the particular learner may be familiar with some functions and contexts for a particular word, but not with the other ones. This poses a substantial problem for the students. We hypothesize that an intelligent tutor can be more adaptive and sophisticated in regard to the semantics of the language if NLP techniques are employed.

Teaching people grammar constructions and vocabulary in Korean with active tailoring to their knowledge level remains an open problem in the area. We, therefore, analyzed the existing state of the field and characterized different aspects of the problem. Our objective was to devise and implement a system for mobile personalized language learning. We propose a prototype solution which is based on the combination of NLP and graph algorithms.

This paper is divided into four parts. The first part gives an overview of the recent related work. We then introduce the problem and outline the main tasks. The third part describes the proposed solution. Finally, we discuss the preliminary results and remaining challenges.

II. Related Work

The Routledge Handbook of Corpus Linguistics [1] provides an overview of the natural language processing used in language learning applications. The review shows that NLP can be successfully used to provide feedback and keep track of the skills acquisition process. Previous studies have mainly focused on error identification and correction. The approaches gradually evolved from analyzing the students response as a whole to extracting specific error patterns. NLP can also be used to automatically extract and annotate a large amount of textual data. In addition, these texts can be indexed and searched in an intelligent manner, by exploiting rich linguistic annotations. What is even more relevant, NLP can be utilized to build reading environments, and generate exercises, which is precisely the focus of our system.

Several studies suggest that existing systems (like flashcards) often lack usage context information [2]. They tend to focus solely on the word without presenting the user with illustrative examples of how they are used in the context. In order to alleviate this issue, Tolmachev *et al.* uses semantic similarity scores and syntactic tree parsing to automatically extract high-quality example sentences [2]. The quality is determined by the sentences being representative of usage patterns and meaning. The authors make an interesting observation that the quality of the example sentences for a learner depends on her current position in a learning process.

Another common problem for language learners is the correct use of collocations. Carlini *et al.* [3] reports that while the language learners use collocations as much as native speakers, their error rate is almost ten times as high. This again brings our attention to the need for the context use examples. The authors propose a search algorithm using an asymmetric normalization of PMI (Pointwise Mutual Information) over a large reference corpus to present the user with the ranked list of corrected suggestions in case a mistake is made.

Kochmar *et al.* [4] addresses the collocation use from a slightly different perspective. The authors note the strong correlation between semantic knowledge and proficiency level. Their experiments suggest that the low-level learners prefer to use high-frequency collocations, most likely because they feel more confident to use commonly established patterns. In the same time, the high-level learners are more adventurous and produce patterns which are still erroneous, but these errors are of a more advanced nature than those of the beginners. The authors adopt a statistical approach and measure the difference between the native and learner language distributions to predict a student's performance.

Ismail *et al.* gives an overview of the existing personalization approaches in the connection with the language learning theories [5]. The paper states that early approaches to foreign language acquisition mainly focused on grammar structures and thus required a long time to obtain a level sufficient for communication, while the modern ones primarily focus on vocabulary acquisition. Therefore, most of the existing personalization systems focus on vocabulary training adjustment. The personalization approaches could include user profiling using questionnaires and activity tracking, user demographics and environment analysis, as well as information on individual cognitive processes. The paper of Petersen *et al.* is an example of user environment personalization, in which vocabulary tutoring is highly connected to the place where a learner accesses the mobile application, and to the learner's declared goals (*i.e.*, museum visit, shopping, *etc.*) [6]. Ismail *et al.* note that the techniques used to adapt the content to a user could be either recommender system like, based on user similarity, or compare features of the content. As one of the prominent works, the paper cites Jung & Graf [7], who used a semantic graph of words associations to model words familiarity. The paper concludes that a major weakness in existing learners assessment approaches is their subjectivity.

Chen *et al.* uses personalization to deal with forgetting during vocabulary acquisition. The proposed system evaluates a learners vocabulary ability based on Item Response Theory, and then recommends proper learning materials. In addition, the learning memory cycle adjusts the review period for learning vocabulary based on an individual learners memory cycle for various words. According to the difficulty of each word and individual learner vocabulary abilities, the proposed system plans adaptively a period for reviewing vocabulary for individual learners [8].

The paper by Řihàk *et al.* is an example of recommender-like personalization approach. The authors measure words complexity by calculation of user performance based on false responses count and on average time spent by a user to produce a correct answer. This measure is then used to cluster tutored items [9].

A recent paper by Wu *et al.* describes a general-purpose neural embedding model StarSpace for embedding of multi-relational graphs, and learning word, sentence or document level embeddings [10]. This work primarily motivates our choice of graph-based representation, as StarSpace method allows us to produce embeddings for a wide range of NLP tasks, as well as recommendation, which would be central for the language learning system.

It should be noted that personalized data-driven language learning systems have rarely been studied directly. A number of questions regarding them remain to be addressed.

III. Problem Statement

To achieve a learner-tailored acquisition of real-world grammar constructs of the Korean language, we propose the following objectives:

- Build a learning corpus of real-world use cases of Korean grammar constructs;
- Build a database of Korean grammar constructs function as a dependency of their context;
- Implement an algorithm to increase a learner's familiarity of context-dependent use cases for each of the grammar constructs.

IV. Proposed Solution

A. System Principle

We propose a semantic graph based personalized tutoring system. It should be noted that within this work, the context is referred to the token context within a sentence, and not to the learner's environment context.

The learning corpus is constructed using real-world sentences from a newspaper, which are translated using an automated service (Google Translate API) and processed with NLP techniques to extract token functions.

A graph is used to track word and grammar construct context and thus find similar and dissimilar word use cases, as well as for the estimation of sentence complexity.

The system then shows words and grammar constructs from real-world sentences to a learner and asks if they understand the word and sentence meaning. If it is not the case, the translation (for tokens and the whole sentence) and grammar

function (for tokens) is provided. The collected context dependent understanding data, its change in time, as well as the sentence complexity estimation are then used to show the learner a language level tailored set of sentences.

B. Learning Corpus

The traditional approach to language learning implies that the learning corpus is constructed around instructional situations with a defined vocabulary, and thus by definition has a limited connection to real-world language. This is due to a huge amount of work required to translate a corpus, define the function of each word in a sentence, and adapt it to the needs of a particular person. We proposed that these limitations could be overcome by using automated services to obtain the translation, and NLP techniques to define the role of each word in a sentence.

We selected 121 articles from The Chosun newspaper[1], published in February 2018, and used the automated downloading procedure to form our private dataset. This dataset contains articles' titles and texts, as well as publication dates. An open source library like "newspaper"[2] or a custom web scraper can be used to extract a larger database.

The articles were further split into sentences and words using regular expressions. The part of speech tagging was carried out using Mecab parser[3] and KoNLPy package[4]. The resultant part of speech tag list covered an extensive set of cases, including basic grammar structures. In addition, we added our own rule-based parser to distinguish between word and grammar tags. The translations of words and texts were acquired via the Google Translate API[5]. One can optionally add openly available online dictionaries (e.g., Naver[6]). The transliterations were obtained with the Hangul Python package[7].

C. Semantic Graph Structure

We proposed that the context-dependent function of a word in various sentences could be represented as a graph, where words and sentences are the nodes, and word functions and contexts are properties of edges between these nodes. One way to capture the context of a word is to connect subsequent words in a sentence on a directional graph. This approach provides a more detailed picture but requires additional techniques to identify similar contexts that define word meaning. At the same time, within the task of teaching Korean context-dependent grammar constructs meaning, the word meaning is usually closely related to its function in a particular sentence. Therefore, we decided to describe the grammar role of each word in the property of an edge to the corresponding sentence, and the user important context in the edges' properties. In addition, we would also like to capture an interdependence of words in a sentence, since it might be help to distinguish different word meanings. For this, we added "Syntactic Dependence" as an additional node type on the graph with directional edges to and from the interdependent words and a non-directional edge to the sentence, in which this dependence occurred. This way, the sentences with similar word dependences could be found by traversing a graph via such nodes.

As far as user input is directly connected to the words and sentences, we decided to include a user as a separate node type on the same graph and record the mentioned input as edges to corresponding words or sentences with a time stamp. Two event types are recorded: "shown to user" and "understanding rating". For the understanding rating of words, the corresponding sentence node id is also stored in the edge property.

The final structure of the cyclic semantic graph built using the data from the learning corpus is presented below (Fig. 1). In addition to the aforementioned nodes and edges, it also has nodes with full texts for advanced learners.

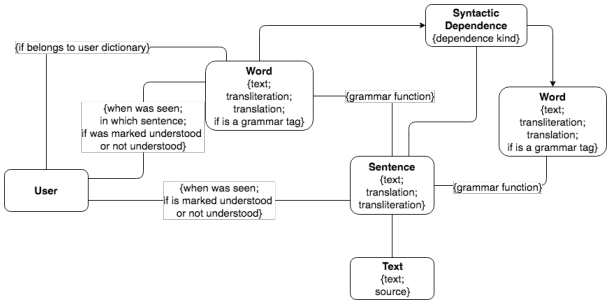

Fig. 1. Semantic graph structure.

D. Semantic Graph Usage

The usage examples for a given word with a given grammar function are obtained by traversing a graph from the word's node. One can add an optional filter by the existence and count of edges to the current user. The sentence complexity is estimated as a number of edges to words multiplied by the number of sentences function types in the edges from each of the connected words. In addition, the familiarity of a learner with words could be estimated by traversing the graph from the user via sentences to words, optionally taking into account the number of "shown to user" edges. In other words, by traversing a graph from a sentence to a user via words, we can estimate sentences complexity as a function of understanding ratings of all words in a sentence and word shown counts with a filter by the word function in a sentence.

When a word is marked by a user as non-understood, other words in a sentence, in which the word was not understood, are used to perform a search of similar settings on the graph.

In future, an extension of the graph structure by adding of a more elaborate syntactic tree of sentences is planned. This will allow for a more precise definition of grammar structure dependent word meanings.

[1]http://www.chosun.com
[2]http://newspaper.readthedocs.io
[3]https://taku910.github.io/mecab
[4]http://konlpy.org/en/v0.4.4
[5]https://cloud.google.com/translate
[6]https://dict.naver.com
[7]https://pypi.org/project/hangul

E. Learning Sequence

The implemented tutoring sequence is based on showing the learner real-world sentences, asking him whether he understands them and the words within them, and querying learning corpus to show more examples of non-understood words. It consists of the following steps.

1) The system shows the learner sentences with grammar constructs in their context.
2) The learner marks the sentences, as well as separate words within them, as understood or not understood, and this rating is stored together with the context.
3) To allow for vocabulary acquisition, meanings of non-understood words are shown to the learner with the ability to build a custom user reference dictionary.
4) The system records the learner activity, that is, which sentences were shown and when which words were shown in the context of which sentence.
5) The words rated not understood are automatically added to the user dictionary, and optionally removed from there after rated understood.
6) The system shows the learner more examples with the non-understood words, and less with understood (based on how many times they were seen and how they rated).
7) The sentence length and complexity are gradually increased based on the accumulated understanding by the learner (average complexity rating of the words marked as understood, average sentence length marked as understood)

In addition, the learner has the ability to read full non-adapted texts with translation tooltips and label the non-understood words and grammar tags from there.

F. System Architecture and User Interface

The system is realized as a web application. The learning corpus and semantic graph are stored in the Neo4j database[8]; the same database is used to store user learning history, profile details and credentials. The web server is implemented using the Flask[9] Python package. A separate Python module carries out the web scraping and NLP analysis task, which were described in detail in the Subsection IV-B, as well as requests to the Google Translate API. The results of NLP analysis and translation are stored in the database for reuse. This module could be run on a regular schedule to update the learning corpus. The web interface is designed based on Reactive principles, allowing for user-friendly scaling and rearrangement of UI elements for smaller device screens. In addition, a separate user interface is implemented for phone screens.

The web interface consists of six pages:

- Login page
- Registration page
- Sentence display page
- Full text reading page
- Pronunciation reference page
- User dictionary page

Each page has a menu with links to all other pages of the system.

The first page, which user encounters, is the login page with the login and password fields, which also has a link to the registration page.

The central page of the desktop version of the web application presents the learner Korean sentences with the list of the words in the sentence with their meanings and role in the sentence (Fig. 2). The screen has buttons for the learner to rate if they understand the sentence as a whole and each particular word within it. If the learner thinks some word interesting or important, he may add it to his dictionary from the same screen.

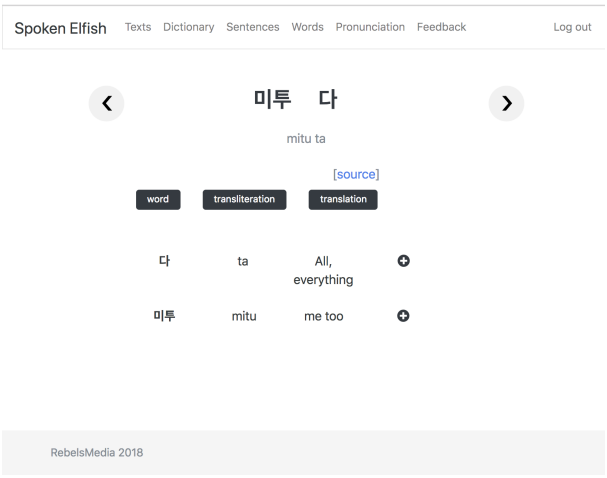

Fig. 2. Desktop web application screenshot.

The mobile version of this page is slightly modified (Fig. 3) to display several sentences at once; it also shows the words of each sentence only on demand. The meaning of each word is displayed only for those rated as ”not understood”.

The dictionary page (Fig. 4) is primarily for learner reference. It contains the words he has marked as not understood, displayed with their meanings and transliterations. Each word has a link to a sentence page with use cases for it. It also has buttons to remove each particular word from the dictionary.

The pronunciation page presents the user a non-interactive reference for Korean alphabet reading rules with examples for each rule.

The full-text reading page is mainly for advanced users. It displays entire news articles. A tooltip with translation and transliteration could be displayed on taping or mouse hovering on each sentence. In addition, separate words could be marked non-understood from here.

V. Conclusion and Future Perspectives

In conclusion, we proposed a system that addresses the problem of training of context-based change of word meaning. We believe this is essential for language learners to become

[8]https://neo4j.com
[9]http://flask.pocoo.org

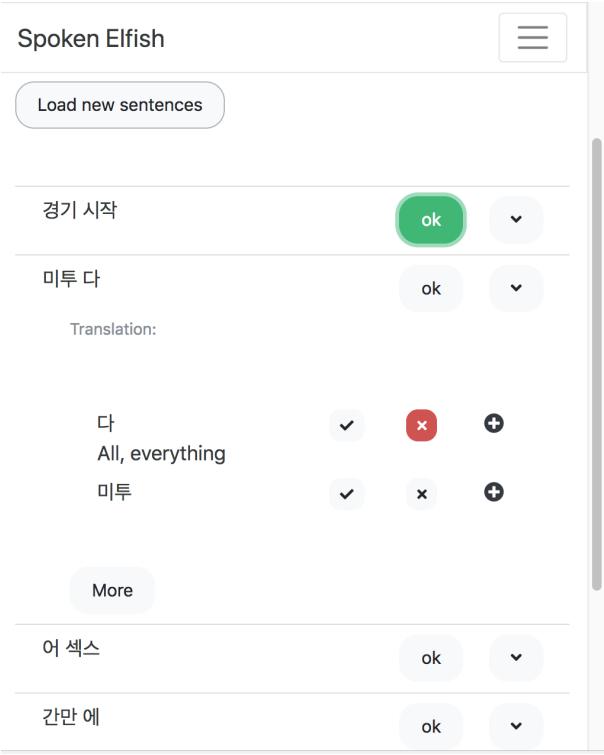

Fig. 3. Mobile web application screenshot.

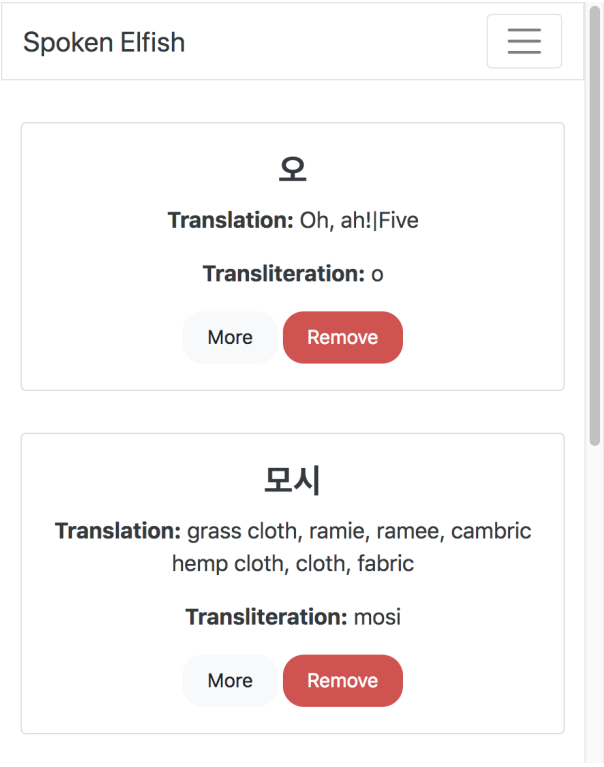

Fig. 4. User dictionary page.

more fluent, however, this is a rather overlooked issue. The usage of a graph with the word contexts and grammar roles captured as its nodes and edges simplified the procedure of presenting a learner with more examples of similar word function and context. Automated translation and NLP were employed to translate sentences, decompose sentence structure and extract word grammar roles. This allowed an easy generation of a large learning corpus of real-world texts and sentences. The system prototype is available online. [10] .

The system was introduced in a class of 30 students. The teacher expressed a general positive opinion and high level of the system acceptance. A follow-up study is currently being planned to assess the efficiency of the system compared to traditional face-to-face tutoring.

The described approach has a potential to improve the performance of personalized tutoring systems for Korean. Our future work will concentrate on enriching semantic annotations of the texts. We plan to use FastText Korean language embeddings[11] to work with the semantics of the language. Our choice is motivated by the high quality of the vectors, as well as their wide recognition. Besides, a follow-up study is currently being planned to thoroughly test the system with the real users' feedback. The preliminary response from a small class of volunteers was overall positive, and we are incorporating teacher's feedback into the system.

Furthermore, we plan to add the possibility to select the sources of texts, and as such we would add more modern expressions, slang and jargon in the vocabulary. In addition, the news data source could be modified for a continuous update, thus making it more interesting for advanced learners. This way, they will have an additional stimulus to use the system as a hybrid for reading news and improving their language.

The tutoring and personalization parts of the system could be further improved by implementing learner performance based word similarity rating [9] and by implementing a more sophisticated review cycle to fully take into account the properties of human short term and long term memory and the individual properties of each learner (e.g. learning frequency, average session duration and learning curve) [8].

The technical side of the system could also benefit from building a mobile application instead of only mobile version of a web application for better user experience and a possibility for learning during unstable Internet access.

To further improve the user experience and motivation, we are going to implement a gamifying element by displaying visualization for daily achievements and weekly/monthly goals.

On a wider level, research is needed to investigate the most common problems that the language learners encounter, as well as their underlying reasons. Furthermore, we do not distinguish between different dialects of Korean as of now, while this can be of interest for the data-driver language learning research community as well. In addition, a directional version

[10] http://www.spokenelfish.com (a registration token is provided upon request)

[11] https://fasttext.cc

of the graph could be used to capture word surroundings and by this capture not only grammar function differences but also a context-dependent translation of words.

VI. Acknowledgment

We thank Aisylu Mulyukova a lot for consultations on Korean grammar (the correctness of translations, accessible explanations of beginner level grammar structures, and a list of examples for the latter).

References

[1] A. O'Keeffe and M. McCarthy, *The Routledge Handbook of Corpus Linguistics*. Routledge, google-Books-ID: giaMAgAAQBAJ.

[2] A. Tolmachev and S. Kurohashi, "Automatic extraction of high-quality example sentences for word learning using a determinantal point process," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 133–142.

[3] R. Carlini, J. Codina-Filba, and L. Wanner, "Improving collocation correction by ranking suggestions using linguistic knowledge," in *Proceedings of the third workshop on NLP for computer-assisted language learning*, pp. 1–12.

[4] E. Kochmar and E. Shutova, "Modelling semantic acquisition in second language learning," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 293–302.

[5] H. M. Ismail, S. Harous, and B. Belkhouche, "Review of personalized language learning systems." IEEE, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7880051/

[6] S. A. Petersen, J.-K. Markiewicz, and S. S. BjRnebekk, "Personalized and contextualized language learning: Choose when, where and what," vol. 04, no. 1, pp. 33–60. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S1793206809000635

[7] J. Jung and S. Graf, "An approach for personalized web-based vocabulary learning through word association games," in *Applications and the Internet, 2008. SAINT 2008. International Symposium on*. IEEE, pp. 325–328.

[8] C. M. Chen and C. J. Chung, "Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle," *Computers and Education*, vol. 51, no. 2, pp. 624–645, 2008.

[9] J. Rihák and R. Pelánek, "Measuring similarity of educational items using data on learners' performance," *Educational Data Mining*, 2017.

[10] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, "StarSpace: Embed all the things!" [Online]. Available: http://arxiv.org/abs/1709.03856