Usability Evaluation of Lexicographic e-Services

Diana Alhafi
Birzeit University
Birzeit, Palestine
diana.alhafi@gmail.com

Anton Deik
Birzeit University
Birzeit, Palestine
anton.deik@gmail.com

Mustafa Jarrar[1]
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

***Abstract*—Although the field of usability evaluation is a well-established discipline, there are no studies on how the usability of lexicographic e-services can be evaluated. This includes, for example, efficiency, effectiveness, and user satisfaction when looking up for synonyms, meanings, or translations using online lexicons. In this paper, we propose to combine two types of usability evaluations to assess the usability of such services: a subjective user-experience evaluation and a more objective controlled experiment—demonstrating how both methods complement each other. We applied our proposed approach to evaluate two important online lexicographic e-services: a lexicographic search engine developed at Birzeit University (https://ontology.birzeit.edu) as well as Google Translate. The user-experience evaluation was conducted through a survey that involved 622 users, and was designed to measure effectiveness, efficiency, satisfaction, and learnability. The controlled experiment involved a set of defined tasks, which were carried out by four teams (12 people) in two laboratories, and their performance was monitored. The tasks were designed to measure effectiveness and efficiency.**

***Keywords*—** *Lexicographic e-Services, Lexicographic Search Engine, Google Translate, Arabic Ontology, Usability Evaluation, User Experience Evaluation, Controlled Experiment.*

I. Introduction and Motivation

Dictionaries are no more limited to the traditional use of hardcopies. Different types of lexicographic services are emerging online [1], ranging from bilingual translation, synonyms, meanings and definitions, semantic differences between terms, spelling, morphology, autocompletion, and more. Lexicography is also becoming a multidisciplinary domain [2], involving new computational methods to derive and build lexicographic data. For example, the Linguistic Linked Open Data Cloud [3] is a community initiative to collect and interlink different lexical resources (130 so far), enabling new lexical information to be derived from this interlinking. Panlex [4] is another ambitious project to collect and integrate about 2500 dictionaries, offering bilingual translation services between many languages.

Although many lexicons are available in digital formats for most languages, there are only a few Arabic portals that provide basic lexicographic services online, such as lisaan.net, albaheth.info, or almaany.com. However, the content of these lexicons is partially structured (i.e., available in flat text), which allows only for basic string-matching searches; e.g., searching for a word would retrieve all the paragraphs that include this word. Furthermore, Arabs mostly rely on using online machine translation tools (e.g., Google Translate) for their language needs, especially term translations. However, such translation tools are not designed for this purpose. Machine translation tools are built on statistical models, and thus, they perform better in translating sentences. Using them to lookup term translations does not yield good accuracy, especially in specialized and domain-specific translations.

A lexicographic search engine (https://ontology.birzeit.edu) was recently developed at Birzeit University [5], allowing people to search for translations, synonyms, definitions, among other lexicographic services – see Fig. 1. The search engine was developed with state-of-the-art design features and according to W3C recommendations and best practices for open data publishing, including the W3C Lemon model [6], which is particularly important for referencing and linking linguistic data. Furthermore, the search engine was built on top of the largest Arabic lexicographic database [2], which comprises about 150 Arabic multilingual lexicons that were manually digitized and then integrated into a normalized database model [7]. The database covers almost all domains, such as natural sciences, technology and engineering, health, economy, art, humanities, and philosophy, among others. It also includes many types of lexicons, such as modern and classical linguistic lexicons, thesauri, glossaries, lexicographic datasets, bi- and tri-lingual dictionaries, as well as the Arabic Ontology – an Arabic WordNet with ontologically cleaned content, used to reference and interlink lexical concepts [8, 9]. The database currently contains about 2.4M multilingual lexical entries, 1.1M lexical concepts, 1.5M translation pairs in Arabic, English and French, 0.7M glosses, and 0.5M semantic relations.

While the growth of digital Arabic lexicographic data and services represents an important milestone in the development of Arabic technologies, there remains a persistent need for these services to be usable. In particular, e-lexicographic services should be effective, efficient, and resulting in user satisfaction.

This paper aims to study the usability of Arabic lexicographic e-services. In particular, we aim to evaluate the usability of Birzeit's lexicographic search engine and to compare it with the usability of Google Translate, which, we assume, is the most commonly used e-lexicographic service among Arabs. To conduct our evaluation, we chose to combine two types of evaluation methods: (i) a *subjective* user experience evaluation, and (ii) a more *objective* controlled experiment. The user experience evaluation was conducted through a survey that involved 622 users, and was designed to measure effectiveness, efficiency, satisfaction and learnability. The controlled experiment involved a set of pre-defined tasks, which were carried out by four teams (12 people) in two laboratories at two universities in Palestine. The tasks were designed to measure the effectiveness and efficiency of the lexicographic search engine in comparison with Google Translate.

[1] Corresponding author

978-1-7281-5052-9/19/$31.00 ©2019 IEEE

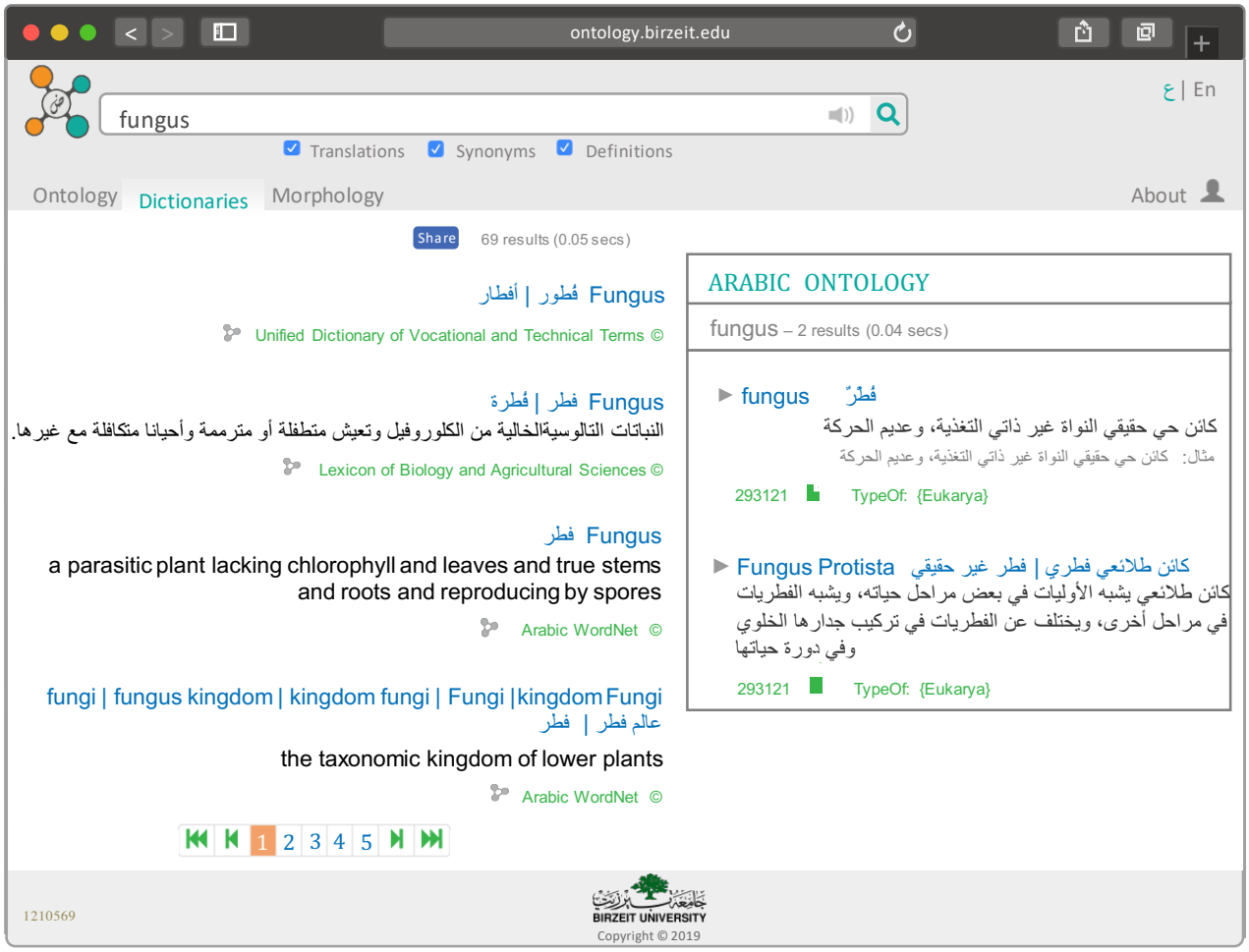

Fig. 1. A Snapshot of our Lexicographic Search Engine

Although the field of usability evaluation is a well-established discipline, it continues to evolve as new technologies and online services are emerging. Up to our knowledge, there are no studies in literature on how the usability of online lexicographic services can be evaluated. Therefore, the originality of our study lies in its attempt to evaluate the usability of a lexicographic service. This is done by combining two different usability evaluation methods: user experience evaluation and controlled experiment. We believe that both methods are important for evaluating a lexicographic search interface, for two reasons. *First*, as will be explained below, both of these methods complement each other as one measures user satisfaction (user experience evaluation) while the other measures task performance (controlled experiment). *Second*, the results of the user experience evaluation, which are subjective, could be confirmed with the more objective controlled experiment. If both the subjective and objective evaluations give similar results, then the experiments' design and their outcomes are realistic; otherwise, one of them might not be well-designed.

The remainder of this paper is organized as follows. Background and related work is presented in section II, the user experience evaluation in section III, and the controlled experiment in section IV. Section V concludes the paper.

II. Background and Related Work

This section presents background and related work in two parts. The first part discusses the definition of usability and its measures, and the second part surveys usability evaluation methods.

A. Usability and its measures

Usability is considered one of the most traditional disciplines in the field of Human-Computer Interaction (HCI). The International Organization for Standardization (ISO) defines usability as a "measure of effectiveness, efficiency and satisfaction with which the product can be used by specified users to achieve certain tasks in a specific context" (ISO 9241-11, 1998). According to this definition, there are a number of criteria or measures used to assess the usability of a piece of software. These include the core three criteria of the ISO definition (effectiveness, efficiency and satisfaction) in addition to other criteria such as learnability and memorability (e.g., see Jakob Nielsen [10]). In this paper, we focus on the three core measures specified by the ISO definition, in addition to learnability. We adopt the definitions of these measures as found in the ISO/IEC 9126-4 standard and in Jakob Nielsen [10], one of the leading experts in the field of HCI. The definitions are as follows:

1- **Effectiveness:** The accuracy and completeness with which users achieve specified goals. In other words, this measure assesses *how accurately the users perform the*

tasks. In our domain, lexicographic tasks include searching for synonyms, looking up translations, finding definitions and meanings, determining semantic differences between terms, and checking exact spelling, among others.

2- **Efficiency:** The resources used in relation to the accuracy and completeness with which users achieve goals. Typical resources include time, human effort, costs and materials. More specific to the purpose and scope of this paper, efficiency assesses *how long it would take the users to perform a lexicographic task*.

3- **Satisfaction:** The comfort and acceptability of use, where the user experience that results from actual use meets the user's needs and expectations. Usually, satisfaction is closely tied to design. Nielsen [10] summarizes this in the following question: *"How pleasant is it to use the design?"*

4- **Learnability:** The easiness by which the user is able to use the system from a first attempt. In other words, *how easy is it for users to perform basic tasks the first time they use the system?*

B. Usability evaluation methods

There are a number of methods in which these four usability measures can be evaluated. These methods can be divided into User-Based Usability Evaluation Methods (UEMs), and methods not involving users. In the latter, the evaluation is done by the researcher without any involvement from users [11]. It can take many forms such as inspections, walkthroughs, modeling, and heuristics, among others. One of the main shortcomings of this method is that the findings are usually not very accurate [11].

User-Based Evaluation Methods are the most commonly used usability evaluation approaches, especially for evaluating websites. These methods can be conducted in a number of ways, such as user testing [11, 12], think-aloud method [13, 14], constructive interaction [15], and eye tracking [16], among others. The think-aloud and the constructive interaction methods are similar. In these methods, users are asked to think aloud during their interaction with a system [13], either alone with a recording device (classical think-aloud approach) or with another user (constructive interaction – which is closer to a natural setting). The idea is to record and understand how users think while interacting with the system. These two methods, however, are more suitable to complex systems and interfaces, rather than to search interfaces as is the case in a lexicographic search engine. In the eye tracking method, special equipment is built into the computer monitor with an eye tracking software that tracks the user's screen. The aim is to record and observe the exact paths the users follow while using the web [16]. While this method is helpful in observing users' behaviors, it fails to detect whether the user is happy or confused when they look at their screen. Furthermore, the special equipment required for this evaluation is usually expensive.

User testing is one the most used user-based evaluation methods. It can be roughly divided into two types. The first type is the ***user experience evaluation***, which uses surveys, questionnaires, and interviews, among others, to measure user experience and satisfaction. This type is usually done in natural settings rather than labs, and it aims to understand how the product will perform in the real world and to study users' behaviors with the new technology [17]. The second type is the so-called ***controlled experiment***, which is usually done in lab settings and aims to measure the typical user performance [11]. Controlled experiments are usually used to evaluate usability measures such as efficiency and effectiveness in a *more objective* way, whereas user experience evaluations focus on *subjective* measures such as users' feelings about the system. In other words, user experience evaluations assess lived experiences, while controlled experiments evaluate task performance [18]. Nevertheless, as convincingly demonstrated by several studies [e.g., 18, 19, 20], these two types of user testing methods are not mutually exclusive, but rather complementary. As rightly noted by [20], controlled experiments aim to improve performance, while the aim of user experience evaluations is to improve user satisfaction.

III. User Experience Evaluation

The first part of our usability study is a subjective user experience evaluation, which was conducted using a survey that involved 622 respondents from different backgrounds. The purpose of the survey was to assess user experience with the four usability measures defined above, i.e., learnability, effectiveness, efficiency, and satisfaction (with design). In what follows, we briefly explain the survey design, data collection method, and respondents' backgrounds. We then present and analyze the results of the survey.

A. Survey Design

The survey (accessible at https://ontology.birzeit.edu/s) was designed using the guidelines suggested by [21]. It consisted of 16 questions. The first four questions (Q1-Q4) collected general information about the respondents: their profession, age, purpose of using the search engine, and frequency of use. Questions Q5-Q16 represented the core of the survey, and were designed to assess the four usability measures explained earlier.

B. Data Collection

About 1000 questionnaires were distributed over a period of 2 months, by publishing the survey on social media channels and mailing lists, and by visiting several universities and directly interacting with employees and students during classes. The total number of valid responses acquired was 622. When it was possible, the evaluator explained the nature of the search engine verbally, in about 1 to 2 minutes, and then asked the respondents to interact with it using as many search words as they wanted. The respondents were then asked to fill in the survey.

C. Respondents' Backgrounds

Figures 2-5 summarize the results of the first 4 questions in the survey. They depict the professions of the respondents (Fig. 2), their age groups (Fig. 3), their purpose of using the search engine (Fig. 4), and the number of searches they made before filling out the survey (Fig. 5). As can be noticed from the

charts, the survey involved people from different professions and age groups, with the majority being students and in the age group 25-30 (see Figs. 2 and 3). This sample of respondents represent the main target audience of the lexicographic search engine.

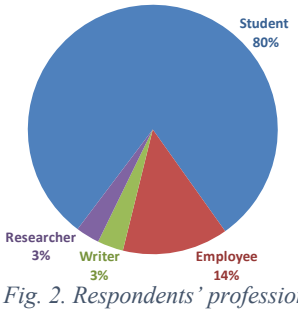

Fig. 2. Respondents' professions

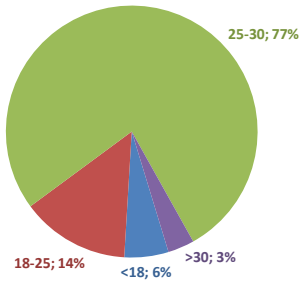

Fig. 3. Respondents' age groups

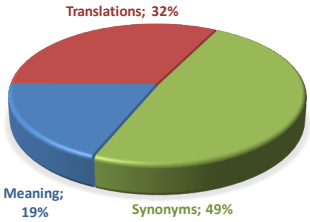

Fig. 4. Purpose of use

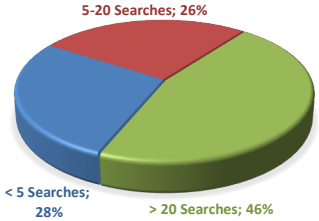

Fig. 5. Frequency of use

D. Survey Results

The 12 core questions in the survey (Q5-Q16) and their results are presented in Fig. 6. All questions were measured using a 4-point Likert scale [19], namely, Excellent, Good, Acceptable, and Weak. The averages were calculated and rated according to the point-based rating system provided by [22].

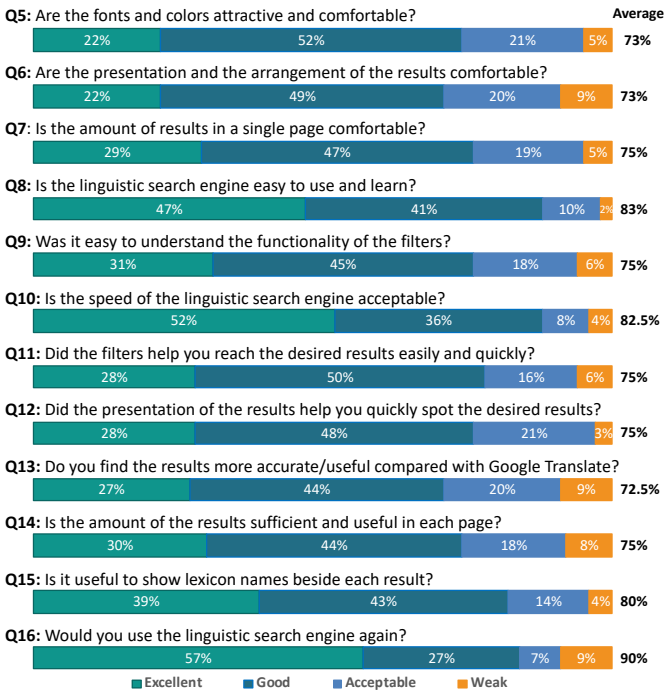

Fig. 6. Results of questions Q5-Q16

We have also studied the effects and correlations of the respondents' backgrounds (questions Q1-Q4) on the results of these 12 core questions. We did this because we wanted to find possible differences, for example, between those who had made less than 5 searches before filling the survey and those who made more, or the differences between young and old users, or between students and none students. The results of these comparisons are not presented here for space limitations but can be found in [23]. Nevertheless, the four variables measured in questions Q1-Q4 bear no significant effect on the results of the 12 core questions (Q5-Q16).

E. Analysis and Discussion

Table I presents a summary of the averages of the survey's evaluation of our four usability measures.

Table I. Averages for each usability measure

Criterion	Satisfaction	Learnability	Efficiency	Effectiveness	Average
Questions	Q5-Q7	Q8-Q9	Q10-Q12	Q13-Q16	Q5-Q16
Average	73%	83%	75%	80%	**77.8%**

Satisfaction (with design). The first three questions (Q5-Q7) assess the users' satisfaction with the design of the search engine, in terms of font and color, presentation and arrangement of the results, and the amount of results displayed in a single page. The average of the responses to these questions is 73%, which indicates a good user satisfaction according to the rating system provided by [22].

Learnability. Questions Q8 and Q9 measure learnability, and are based on the users' first interaction with the system. The questions assess how easy it is to use the search engine and understand the functionality of the filters. Their average is 83%, indicating excellent user experience with regards to learnability, according to [22].

Efficiency. Questions Q10-Q12 measure efficiency in terms of the speed of the search engine and how the filters and the presentation of the results help the users spot the desired results quickly. Their average of 75% indicates good user experience [22].

Effectiveness. The last four questions (Q13-Q16) measure effectiveness. They assess the accuracy of the results, their usefulness, and their sufficiency. Their average of 80% indicates excellent user experience with regards to the effectiveness of the search engine [22].

It is worth noting that most of our usability measures have an average of 75% or more, with the exception of users' satisfaction with design, which had an average of 73%. The reason behind this, according to respondents' comments on the survey, is two-fold. First, many users found that using the color blue to display the results was misleading, as it gave them the impression that the results were hyperlinked, while in fact they were not. Second, many users expressed that the "ontology box" on the right side of the page was confusing and they did not understand the purpose of it. The overall average of the 12 core questions is 77.8%, which indicates a good overall user experience with the usability of the lexicographic search engine. As will be discussed in the next section, this outcome

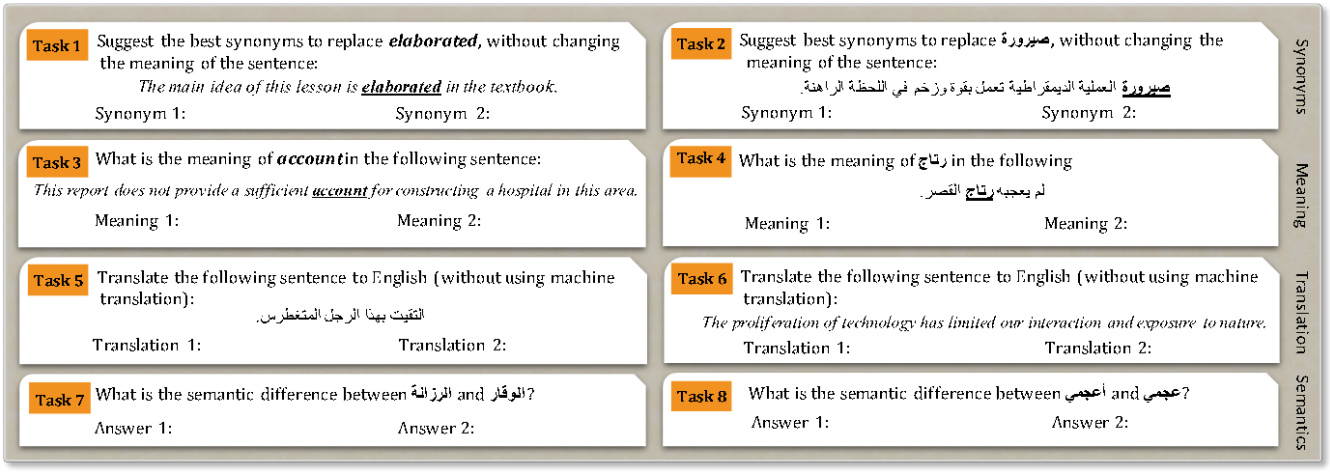

Fig. 7. The eight tasks of the controlled experiment

was also validated in a more objective experiment that yielded similar results. At the same time, we took into consideration respondents' feedback that resulted in a relatively low average for satisfaction with design. In particular, we limited the use of blue color to hyperlinked results and decided to provide more explanation about the purpose of the ontology box at the right-hand side of the page.

IV. Controlled Experiment

Although the evaluation above gives a good overview on the usability of the lexicographic search engine, the study primarily measures user experience. As mentioned earlier, such evaluations tend to be subjective as they depend on the personal taste and experience of the surveyed users. In order to obtain a more objective evaluation, we conducted a controlled experiment to assess two key usability measures, namely, efficiency and effectiveness, for both Google Translate and the lexicographic search engine.

A. Experiment Setup and Design

A set of eight tasks were designed to be carried out in a controlled environment (see Fig. 7). The tasks were designed to measure the efficiency and effectiveness of both Google Translate and the lexicographic search engine by asking the participants to search for (*i*) synonyms, (*ii*) meanings, (*iii*) translation, and (*iv*) semantic differences between terms. These four types of tasks were designed to be carried out both English-to-Arabic and Arabic-to-English, resulting in 8 different tasks. We believe these tasks cover the most common lexicographic services people need, as we learned from interactions with respondents in the user experience survey.

Two controlled experiments were carried out at two different universities; Birzeit University and Palestine Technical University – Kadoorei. The experiments were conducted in a lab environment using Windows PCs. Two groups (A and B) were formed for each experiment at each university; each group consisted of 3 students (12 participants in total). Group A was assigned to use our lexicographic search engine and group B was assigned Google Translate. Both groups were given 10 minutes before the experiment in order to try both tools and get familiar with them.

The tasks were distributed to each participant as a hardcopy, and each task was explained to all the participants before the start of the experiment. In addition, we allowed the participants to provide two answers for each task, such that the participants were asked to record only one answer for the task in case they are very confident in their answer. Otherwise, they can provide a second answer. This is important to let us know how confident the participants were about the correctness of their answers, as will be discussed below. The time needed to accomplish each task was carefully tracked for each of the participants, and their interaction within the group was observed to record any errors.

B. Efficiency

The efficiency of the lexicographic search engine (LSE) was measured vis-à-vis Google Translate (GT) using the metric "task time" [24], which measures the time it took each participant to complete each of the eight assigned tasks. Table II records the detailed results of the two experiments conducted at Birzeit University (BZU) and Palestine Technical University – Kadoorei (PTUK), for all the participants (P1-P12). In addition, Table II records the meantime of each task, calculated based on [24, 25]. The meantime is also represented graphically in Fig. 8.

C. Effectiveness

Effectiveness was measured using the metric "task completion" [24], which assesses the accuracy of the results. This was accomplished by evaluating the two answers that each participant provided for each task against a predefined answer key. As explained earlier, participants were asked to provide only one answer if they were certain about it, and not to provide a second answer unless they think it is also correct in the provided context. A score was given to each task as follows:

- If one or two correct answers were provided, the task was given a score of 10.
- If one answer was correct and the second was a *close* answer (i.e., correct in a related context), the task was given a score of 8.
- If one answer was correct and the second was *incorrect* (or correct in a different context), the task was given a score of 6.

Table II. Task Time (in seconds) by participant by task

Participants			Tasks								
			T1	T2	T3	T4	T5	T6	T7	T8	Mean time
Group A: LSE	BZU	P1	58	27	40	20	45	80	116	95	**60**
		P2	43	17	62	30	53	50	116	118	**61**
		P3	30	24	37	51	50	113	115	96	**65**
	PTUK	P4	110	39	60	31	25	195	123	85	**84**
		P5	116	48	77	43	27	190	120	77	**87**
		P6	50	40	67	43	37	160	150	102	**81**
Meantime			**68**	**33**	**57**	**36**	**40**	**131**	**123**	**96**	**73**
Group B: GT	BZU	P7	90	25	50	30	61	85	60	70	**59**
		P8	120	14	120	61	50	128	60	30	**73**
		P9	110	30	80	60	71	220	60	30	**83**
	PTUK	P10	27	26	40	15	18	75	11	12	**28**
		P11	37	27	43	11	22	85	17	30	**34**
		P12	14	20	44	30	30	90	30	11	**34**
Meantime			**66**	**24**	**63**	**35**	**42**	**114**	**40**	**31**	**52**

Table III. Task Score by participant by task

Participants			Tasks								
			T1	T2	T3	T4	T5	T6	T7	T8	Avg
Group A: LSE	BZU	P1	8	10	10	10	10	8	10	10	**9.50**
		P2	6	10	4	8	10	6	6	8	**7.25**
		P3	10	10	10	10	6	8	6	6	**8.25**
	PTUK	P4	10	10	0	10	8	6	8	8	**7.50**
		P5	8	10	0	6	4	10	6	10	**6.75**
		P6	8	10	8	8	8	8	6	8	**8.00**
Average			**8.33**	**10**	**5.66**	**8.66**	**7.66**	**7.66**	**7.00**	**8.33**	**7.91**
Group B: GT	BZU	P7	0	0	4	4	10	8	0	0	**3.25**
		P8	8	0	4	0	6	10	0	0	**3.50**
		P9	8	0	10	0	8	8	0	0	**4.25**
	PTUK	P10	8	0	2	0	6	8	0	0	**3.00**
		P11	4	0	4	4	6	4	0	0	**2.75**
		P12	10	0	4	0	4	8	0	0	**3.25**
Average			**6.33**	**6.67**	**4.66**	**4**	**6.66**	**7.66**	**0**	**0**	**4.5**

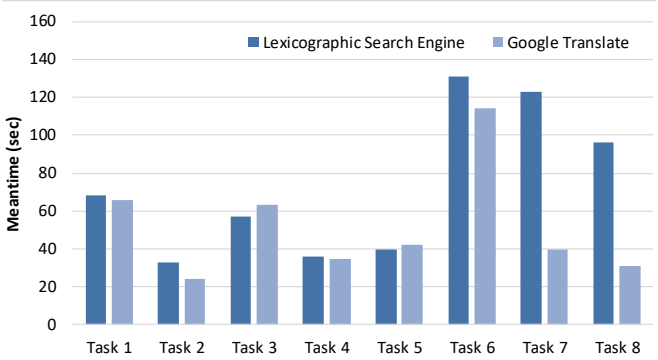

Fig. 8. Meantime, in seconds, for each task (i.e., efficiency)

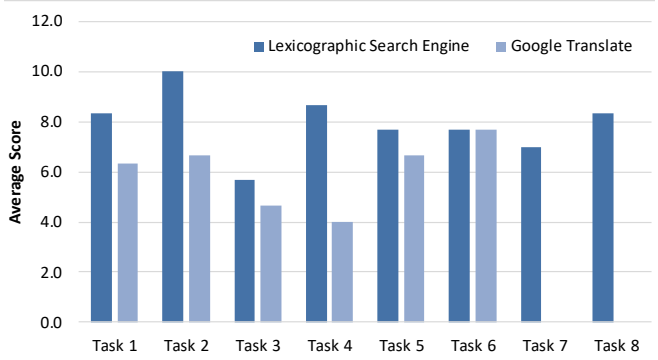

Fig. 9. Average task scores for both tools (i.e., effectiveness)

- If one or two *close* answers were provided, the task was given a score of 4.
- If one answer was *close* and the second was *incorrect* (or correct in a different context) the task was given a score of 2.
- Otherwise, the task was given a score of 0.

Table III presents the scores of each task for all the participants along with the average score of each task. Average task scores are also depicted graphically in Fig. 9.

D. Analysis and Discussion

Our efficiency evaluation (Table II and Fig. 8) shows that Google Translate was a little faster than the lexicographic search engine, (52 sec) versus (73 sec) respectively. In tasks T1-T6, the participants were able to retrieve the answers in both tools in almost the same time. However, in T7 and T8, Google Translate was significantly faster because it did not provide any results (see Table II).

With regard to effectiveness (see Table III and Fig. 9), the results of the lexicographic search engine scored higher than Google Translate in all tasks. In tasks T3 and T4, the results of Google Translate were low, but the worst cases were in T7 and T8 as it was unable to answer any of them. This means that Google Translate is not very effective if used to lookup meanings and semantic differences between terms. In general, the total average of the effectiveness of the lexicographic search engine for all tasks is (7.91/10), while the average of Google Translate is (4.5/10).

The results of the different types of evaluations (the user experience and controlled experiment evaluations) were indeed close to each other. The user experience evaluation results indicated a good overall user satisfaction with the lexicographic search engine, especially that the efficiency was 75% and the effectiveness was 77.9%. The controlled experiment also demonstrated similar results: the effectiveness averaged 7.91/10 and the efficiency 73 seconds. In other words, the 620 survey respondents believe, subjectively, that the lexical search engine is 77.9% effective. In the controlled evaluation, the results of the 12 participants scored a more objective 7.91/10. Similarly, the 75% efficiency in the subjective evaluation and the 73 seconds in the more objective evaluation also demonstrate reasonably-matching results.

We believe that obtaining similar results using two different evaluation methods implies that: (*i*) both methods complement and confirm the results of each other, indicating the soundness of the results, and (*ii*) the experiment design in both methods was indeed realistic in evaluating lexicographic e-services.

V. Conclusions and Future Work

This paper proposed to use two different usability evaluation methods to evaluate the usability of lexicographic

e-services, specifically in searching for synonyms, meanings, translation, and semantic differences between terms. The first is a subjective user experience evaluation that assesses the overall user satisfaction, and the second is a more objective controlled experiment that evaluates task performance. The two methods were used to evaluate the usability of Birzeit's lexicographic search engine and Google Translate. The user experience evaluation was conducted using a survey that involved 622 participants and was designed to measure the effectiveness, efficiency, satisfaction (with design) and learnability. The controlled experiment involved a set of defined tasks, which were carried out by four teams (12 people) in two laboratories at two different universities in Palestine. The tasks were designed to measure the effectiveness and efficiency of the lexicographic search engine in comparison with Google Translate. We plan to extend our evaluation to include other lexicographic e-services, such spell checking, diacritic checking, autocompletion, and others. We also plan to evaluate the usability of other non-Arabic lexicographic tools and compare them with the tools used in this paper.

ACKNOWLEDGMENTS

We wish to thank all people who participated in the survey and the controlled experiment. We are also thankful to Hamzeh Amayreh for his technical support.

REFERENCES

[1] S. Granger and M. Paquot, eds., Electronic Lexicography, Oxford University Press, 2012.

[2] M. Jarrar and H. Amayreh, "An Arabic-Multilingual Database with a Lexicographic Search Engine," Proceedings of the 24th International Conference on Applications of Natural Language to Information Systems (NLDB), 2019.

[3] J. P. McCrae, C. Chiarcos, F. Bond, P. Cimiano, et al, "The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud," The International Conference on Language Resources and Evaluation (LREC), 2016.

[4] D. Kamholz, J. Pool, and S. M. Colowick, "PanLex: Building a Resource for Panlingual Lexical Translation," The International Conference on Language Resources and Evaluation (LREC), 2014.

[5] M. Jarrar, "Search Engine for Arabic Lexicons," Proceedings of the 5th Conference on Translation and the Problematics of Cross-cultural Understanding, The Forum for Arab and International Relations, Qatar, 2018.

[6] M. Jarrar, H. Amayreh, and J. McCarae, "Representing Arabic Lexicons in Lemon - a Preliminary Study," The 2nd Conference on Language, Data and Knowledge (LDK), Germany, 2019.

[7] H. Amayreh, M. Dwaikat, and M. Jarrar, "Lexicon Digitization - A Framework for Structuring, Normalizing and Cleaning Lexical Entries", Technical Report, Birzeit University, 2019.

[8] M. Jarrar, "The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content," Applied Ontology Journal, 2019 (Forthcoming).

[9] M. Jarrar, "Building a Formal Arabic Ontology," In Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks, ALECSO, Arab League, 2011.

[10] J. Nielsen, "Usability 101: Introduction to Usability," http://www.nngroup.com/articles/usability-101-introduction-to-usability/ (Accessed, June 2019).

[11] J. Preece, H. Sharp, and Y. Rogers, Interaction Design: Beyond Human-Computer Interaction, Hoboken: Wiley, 2015.

[12] J. S. Dumas and J. C. Redish, A Practical Guide to Usability Testing, Portland: Intellect, Revised Edition, 1999.

[13] J. Lazar, Web Usability: A User-Centered Design Approach, London: Pearson Education, 2006.

[14] M. Van den Haak and M. de Jong, "Analyzing the Interaction between Facilitator and Participants in Two Variants of the Think-Aloud Method," In the Proceedings of IEEE International Professional Communication Conference, 323-327, 2005.

[15] A. Holzinger, "Usability Engineering Methods for Software Developers," Communications of the ACM, 48(1), 92-99, 2005.

[16] J. Nielsen and K. Pernice, Eye Tracking Web Usability, San Francisco: New Riders Press, 2010.

[17] F. Jambon and B. Meillon, "User experience evaluation in the wild," In CHI'09 Extended Abstracts on Human Factors in Computing Systems (pp. 4069-4074), ACM, 2009.

[18] A. P. Vermeeren, E. L. C. Law, V. Roto, M. Obrist, J. Hoonhout and K. Väänänen-Vainio-Mattila, "User experience evaluation methods: current state and development needs," In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 521-530, ACM, 2010.

[19] K. Finstad, "The Usability Metric for User Experience," Interacting with Computers Journal, 22(5), pp. 323-327, 2010.

[20] N. Bevan, "What is the difference between the purpose of usability and user experience evaluation methods?" In Proceedings of the Workshop UXEM, Vol. 9, pp. 1-4, 2009.

[21] E. Martin, "Survey Questionnaire Construction," Research Report Series, Survey Methodology #2006-13, U.S. Census Bureau, Washington D.C., 2006.

[22] S. H. Mustafa, and L. F. Al-Zoua'bi, "Usability of the academic websites of Jordan's universities an evaluation study," In Proceedings of the 9th International Arab Conference for Information Technology, pp. 31-40, 2008.

[23] D. Alhafi and M. Jarrar (supervisor): "User Experience Evaluation For Birzeit Linguistic Search Engine," Seminar Report, Birzeit University, 2019, http://www.jarrar.info/publications/DianaAlhafiReport11.pdf (Accessed, June 2019).

[24] F. J. García, M. Lozano, F. Montero, J. A. Gallud, P. González, and C. Lorenzo, "A controlled experiment for measuring the usability of webapps using patterns," In Enterprise Information Systems VII, pp. 257-264, 2007.

[25] X. Yuan, H. Yang, K. Moorhead, and K. DeMers, "Evaluating an Education Department Portal: A Case Study," In International Conference of Design, User Experience, and Usability, pp. 240-247. 2015.