Crowdsourcing for building and maintaining multilingual knowledge resources

Mercedes Huertas-Migueláñez
Department of Information Engineering and Computer Science
University of Trento
Trento, Italy
mdlm.huertas@unitn.it

Abstract—Existing lexical-semantic resources are costly in terms of building and maintenance and might be incomplete concerning the knowledge included. Manual development is expensive although it produces high quality knowledge. The overall aim of our research is to build high-quality resources that can be used for different kinds of semantic services such as meaning extraction, data integration and linking, semantic search, and semantic visualization. We propose crowdsourcing as a solution to build and maintain multilingual lexical-semantic resources. In building the resources we differentiate between language-independent and language-specific levels. We have designed a framework for the localization of lexical-semantic resources. In order to evaluate and refine it, we are conducting user studies. In this paper, we describe our framework and the results of the preliminary user study performed on expert users. We plan to gradually involve common users (the crowd) in the localization process, that is, in the production of the lexical-semantic resource in different languages.

Keywords—crowdsourcing; multilingual lexical-semantic resource; user study

I. INTRODUCTION

The varying quality of existing lexical-semantic resources such as WordNet [1] influences the quality of services and applications that use them, such as meaning extraction, data integration and linking, or semantic search applications. There are a number of efforts to build such resources in different languages as in, e.g., [2]–[5]. While these efforts demonstrate the feasibility of manual translation on such resources, various issues remain concerning the quality of the knowledge created in this way: incompleteness, redundancy and ambiguity hamper the usage of lexical semantic resources in applications such as data integration and linking or meaning extraction.

Lexical-semantic resources can be built using different techniques. Manual development has demonstrated to be expensive in terms of required human power, but highly valuable with respect to the quality of the resource produced as claimed in [6]. Human computation, for instance crowdsourcing and gamification, has been extensively used in translation projects such as in [7]–[11]. While crowdsourcing involves great amount of humans to solve different tasks, gamification uses games to get humans solve tasks that computers can't do [12].

The *Universal Knowledge Core* (UKC) is a knowledge base developed at the University of Trento. It is designed as a multilayered ontology that has a language-independent (semantic) and a language-specific (lexical-semantic) layer. In this respect, it provides mappings of common lexical elements from different languages to formal concepts. The UKC is verticalized along language-specific vocabularies. At the moment, vocabularies exist at varied levels of completion for nine languages[1]. This verticalization can't be fully automated and requires manual effort. In order to implement this process, we have designed a step-by-step language development workflow that leverages on human knowledge and skills.

This paper establishes a software framework for expert- and crowd-based building and maintenance of multilingual lexical-semantic resources. This software framework implements the development workflow. This development workflow includes:

- free translation from existing resources, e.g., of English glosses;
- lexicalization by adding lexical or semantic elements form scratch;
- reuse of existing resources, e.g., automated conversion of an existing WordNet-like resource;
- validation of lexical or semantic elements.

The goal of the framework is to accelerate the development process while providing quality control mechanisms over the input through the design of interactive interfaces.

The paper is organized as follows. Section II gives a brief overview of the state of the art. Section III describes the UKC architecture. Section IV presents our approach for crowdsourcing language development workflow. Section V describes the current expert-based system that we have implemented and the results of validation. Future work is presented in section VI.

II. STATE OF THE ART

Our work is situated in a larger area of crowdsourced translation projects.

BabelNet [13] is a large multilingual semantic network created by automatically merging WordNet senses and Wikipedia entries. For the missing lexicalizations BabelNet uses

[1] The languages included are: Italian, Chinese, Mongolian, Hindi, Bengali, Spanish, Catalan, Galician and Basque.

Google Translate API[2] to translate sentences from English into specific languages. After that, they create a set with the most frequent translations for a specific term. Expert annotators validated translations in five languages: Catalan, French, German, Italian and Spanish. The BabelNet team has developed their own video games [14] for the validation of semantic relations and sense-image mappings. The manual validation of the elements included in BabelNet covers a very small proportion of the total suggesting that the resource may include incorrect content.

Duolingo[3] [15] is a website that leverages on gamification where users can learn languages. The underlying aim of Duolingo is to have common people to translate text coming from different sources since machine translation is not good enough and professional translators are costly as stated in [16]. Duolingo helps people to learn a foreign language by asking them to translate sentences from that language to English, starting with simple sentences and advancing to more complex ones as user's skills increase. The game provides one-to-one dictionary translations of individual words, but players use their experience to see what makes sense in context. Duolingo is mainly focused on translation rather than building a multilingual lexical-semantic resource. The crowd to which Duolingo is aimed is willing to learn a language while the crowd needed to build and maintain our resource has to be bilingual and moved by other type of motivation such as satisfaction to contribute to the community.

Wikipedia[4] is a well-known online encyclopaedia whose content is built and maintained by crowdsourcing. It offers the content in 284 different languages. In spite of the efforts to validate the huge amount of existing articles some errors can be found in localized articles. Also due to disagreement of the authors some inconsistences might appear [10].

III. UKC ARCHITECTURE

The UKC is a multilingual knowledge base that contains a layer of language-independent ontology of concepts called *concept core*, connected to a set of language-dependent *vocabularies*, as shown in Fig. 1. The UKC includes other layers such as the entity core, the natural language core and the domain core. For the shake of clarity they will not be described here since they are out of the scope of this paper.

Just like in WordNet, a vocabulary consists of word forms, lemmas, senses, and synsets, sets of synonymous words, with the addition of the notion of *lexical gap* that denotes missing lexicalizations in a given language. The concept layer thus acts as a hub for the lexicalization of concepts in different languages.

A *Local Knowledge Core* (LKC) can be seen as a working copy from the UKC of the concept core and two vocabularies: English and its corresponding localization. The process of localization is executed in the LKC. If new concepts are locally created, they will be added to the UKC during the synchronization process.

Developing a multilingual lexical-semantic resource such as the UKC is a challenging task that requires a scalable software framework to allow the development of the vocabularies and ensures high quality of the resource.

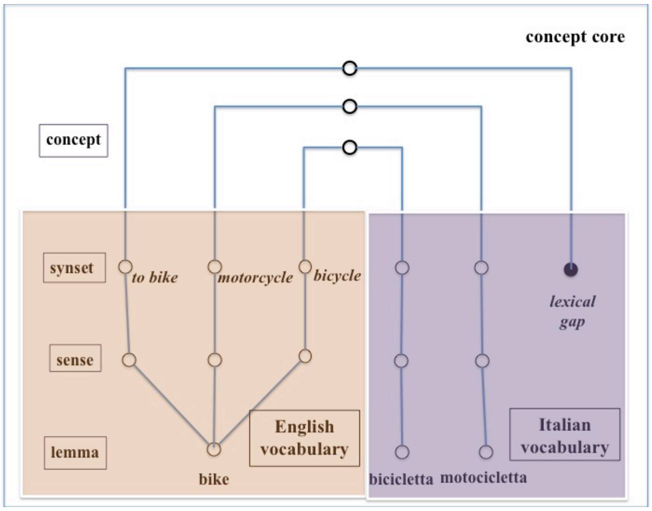


Fig. 1. Sample of the concept core and two vocabularies where there is no word in Italian to express the action of biking so it will be represented on the synset level as a *lexical gap*.

IV. CROWDSOURCING DEVELOPMENT WORKFLOW

Our research concerns the development of a high-quality multilingual lexical-semantic resource, the UKC, in a scalable manner mainly through human effort. In order to support scalability while maintaining high quality we make use of the following techniques:

- reuse of existing resources where allowed by their licenses;
- expert-based manual translation from resources in other languages;
- crowdsourcing in order to accelerate development.

These techniques are integrated in the framework presented in Fig. 2. The novelty of the framework proposed relies on: 1) the UKC that allows interoperability among the languages; 2) the design of a step-by-step language development workflow to ensure the quality of the content of the UKC during the language development process. Quality is guaranteed thanks to the definition of three user roles. Every registered user will have one role assigned and they will receive assignments, sets of tasks to be completed via the *User Interface* (UI). These roles are:

- *LKC developer* is the main contributor who builds the target resource (e.g., by translating from a source or by providing lexicalizations from scratch);
- *LKC validator* evaluates the translations produced by the developers and the new terms they have introduced;
- *LKC manager* is in charge of evaluating if validators have assessed correctly developer's work. The *LKC*

[2] http://research.google.com/university/translate/ Last accessed: 04-August-2015

[3] https://www.duolingo.com/ Last accessed: 04-August-2015.

[4] http://www.wikipedia.org/. Last accessed: 04-August-2015.

manager is also in charge of synchronizing the LKC with the UKC by merging their contents.

The quality of user's contributions will be also guaranteed by measuring their performance. The performance will represent the reputation of a user in the system. Initially, performance will measure the number of tasks completed per time unit and after, we will improve the algorithm to also include the quality of the result of every task. The result of this algorithm will represent the reputation of the user. The reputation values will be higher if the results produced have high quality.

3) The gradual involvement of crowd users by the development of the three systems that compose the framework.

- *System 1* uses experts in the role of *LKC developers* and *LKC validators*. An expert is a person whose mother tongue is the same as the vocabulary under development, has a competent level of English and is, at least, Bachelor student. They will have to complete different assignments according to their role.
- *System 2* engages crowds in the language development process. Crowd users will assigned the *LKC developer* role. On the other hand experts, in the role of *LKC validators* ensure that the assignments are properly completed. In this system the first incentives will be applied to engage users in the development. Also a preliminary reputation algorithm will be used to obtain insights on the performance of users.
- *System 3* engages crowd users both in the development and in the validation process. Incentives will be improved and new models can be created understanding which pair incentive-task produced the desirable results in terms of human satisfaction and task performance.

Systems 1 to 3, implemented in this order, involve the crowd gradually. The sequential order of the creation of these systems will help us to foresee the requirements in their development as well as to understand what users would expect or need when working with a framework like this. We plan to run several user studies along the development of the framework to evaluate such systems and with the results obtained from these studies, we will improve the current system and move progressively to the next one by the addition of new features.

V. CURRENT IMPLEMENTATION

Here we report, in a detailed manner, the architecture of System 1, in which expert developers complete specific vocabulary development tasks and expert validators assess their performance.

A. System Description

As shown in Fig. 2, initially the LKCs can be created by automatically importing an existing WordNet project. The import is performed using a Java programme that takes the original files and creates new ones with the format required to create the corresponding resource. Also, an LKC can be created by relying on experts who create it by translating lexical items from the English WordNet into the language to be developed. Development tasks are completed through a dedicated UI. In Fig. 3 we present the interface for validation where an expert validator accepts or rejects translations made by an expert developer. On the left hand side of the image is presented the

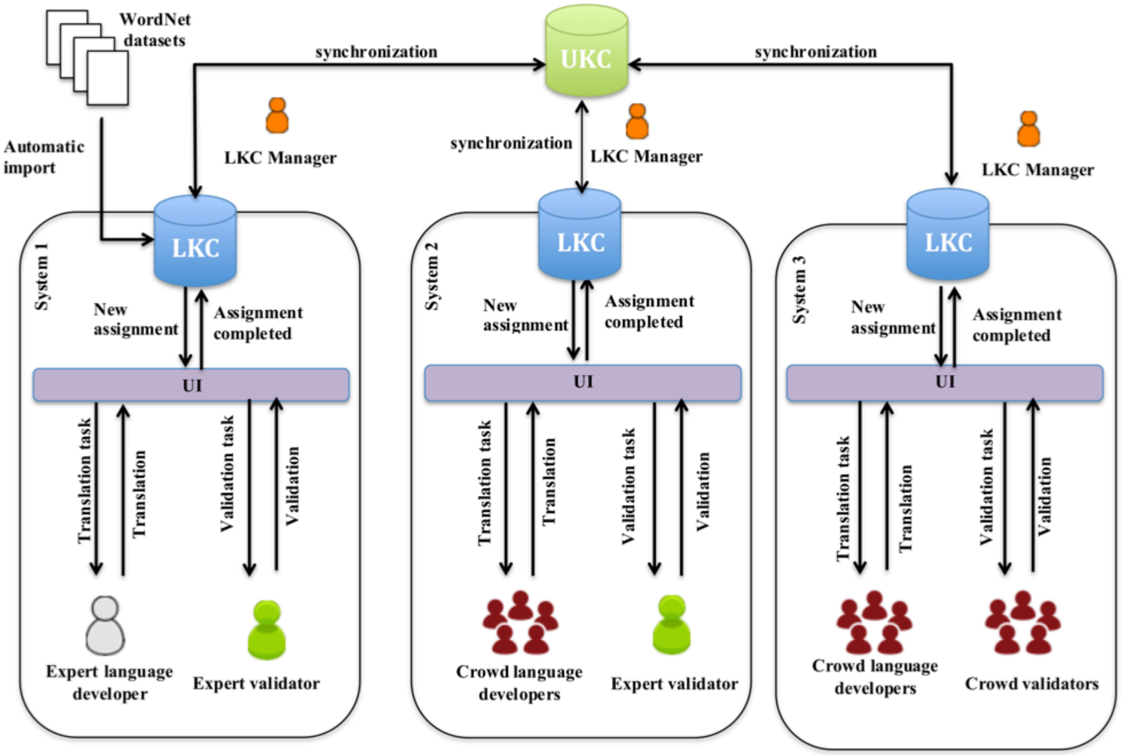
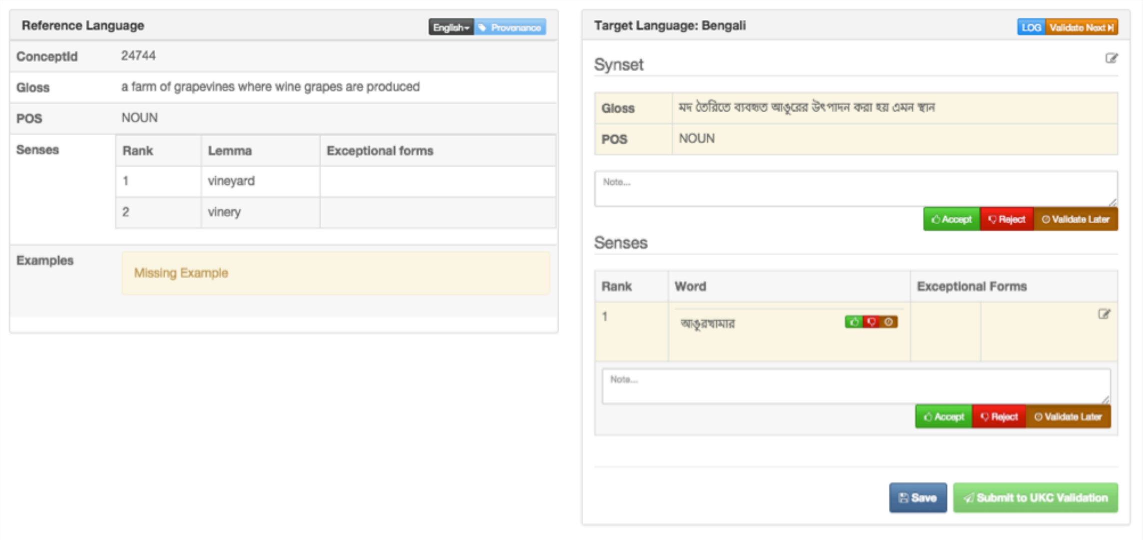


Fig. 2. Framework for expert and crowd based building and maintenance of multilingual lexical-semantic resources. The UKC allows language interoperability. Every LKC allows the development of a specific language. Users develop the LKC via the UI by completing assignments.



Reference Language

ConceptId	24744
Gloss	a farm of grapevines where wine grapes are produced
POS	NOUN

Senses

Rank	Lemma	Exceptional forms
1	vineyard	
2	vinery	

Examples: Missing Example

Target Language: Bengali

Synset

Gloss	মদ তৈরিতে ব্যবহৃত আঙুরের উৎপাদন করা হয় এমন স্থান
POS	NOUN

Note...

Senses

Rank	Word	Exceptional Forms
1	আঙুরখামার	

Note...

Fig. 3. The LKC validator view of the concept *'vineyard'* translated into Bengali. The translation can be accepted or rejected.

task in English, the reference language. On the right hand side, the expert validator can see how the expert developer has lexicalized it.

B. *System evaluation*

We have run a small user study to assess the functionality of the UI in System 1 of the proposed framework and also to better understand users' needs and expectations.

1) *Experiment design*

Participants were professors and students from Bachelors to PhD from universities in China, Mongolia and Bangladesh. The total number of participants was 15, male and female. Their age ranged from 21 to 50. The experiment run from December 2014 to February 2015. The distribution of the roles was made according to the level of education of the participants: those holding a PhD were assigned an expert validator role and the rest were assigned expert developer role.

Once the participants got the role assigned they were asked to perform sets of assignments. After, they were asked to complete a questionnaire. The questionnaire contained three main parts:

- background information, to give general information about the user;
- usability common questions, they were the same independently of the role;
- role specific usability questions, they were dependent of the role of the user.

Questionnaires included two types of questions:

- close questions based on the Likert scale as presented in [17];
- open questions where the responses could be free text.

Our analysis will be focused on the second part of the questionnaire, the usability common questions, where the results obtained were more interesting. This part of the questionnaire had twelve questions that measured eight different usability dimensions[5]. Some of them measured a specific aspect of the UI and some others were counter questions to observe if the user was consistent when responding.

Here we describe the results of the study for the role of LKC developer. The questionnaires corresponding to other roles did not offer interesting results, therefore they are not considered in this study.

2) *Results*

The results presented in Table 1 correspond to the usability common questionnaire. Questions two and seven have high values for the standard deviation suggesting that users had very different opinions on the same usability question. As a consequence, we will change the workflow of the system, improve the UI appearance and the way the system recovers when it blocks, meaning that certain functionalities need to be redesigned. Besides, question three, which is the counter question of question seven, has high value for the standard deviation, which suggests that the way the UI works should be improved.

In the open questions, where users could introduce suggestions and comments, some users suggested that the visualization of the task progress should be present while logged-in in the system, "*The list for the words individuals have done would be very informative*" and that the navigation could be improved to have a "*More user friendly navigation system*".

After analyzing these preliminary results we have defined the list of changes to be introduced:

- improvement of the workflow of the system by redesigning some functionalities;
- improvement of the UI appearance by adding the visualization of the progress done by a specific user and the addition of concept visualization.

[5] http://www.nngroup.com/articles/usability-101-introduction-to-usability/

TABLE 1 RESPONSES OF DEVELOPERS TO COMMON USABILITY QUESTIONS. THE TABLE ALSO SHOWS THE DIFFERENT USABILITY DIMENSIONS TO BE MEASURED FOR EVERY QUESTION

Usability dimension	Usefulness	Error recovery	Satisfaction	Intuitiveness	Memorability	Learnability	Satisfaction (counter question)	Visibility of system's status	System's feedback quality	Learnability (counter question)	Efficiency	Satisfaction
user	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**
1	4	3	3	4	4	4	2	4	4	2	4	4
2	3	2	1	2	3	3	5	3	3	3	3	2
3	4	3	4	4	5	4	2	5	3	3	4	4
4	5	5	5	4	5	5	5	5	5	1	5	5
5	5	4	4	4	5	5	1	5	4	4	4	4
6	4	4	4	4	5	5	1	4	4	4	5	5
7	4	4	4	4	4	4	4	4	4	4	4	4
8	4	1	4	2	5	4	4	4	4	4	4	3
9	3	4	4	2	5	5	2	4	4	4	5	5
10	5	5	4	4	5	4	2	5	5	3	4	4
11	5	4	4	4	5	4	2	4	4	4	4	4
Mean	4.181	3.545	3.727	3.454	4.636	4.272	2.727	4.272	4	3.272	4.181	4
S.D.	0.715	1.157	0.962	0.890	0.642	0.616	1.420	0.616	0.603	0.962	0.574	0.852

Q1 - I find the system simple to use.
Q2 - In case I make a mistake, I can recover easily and quickly.
Q3 - The system has all the functions I expect it to have.
Q4 - I do not notice any inconsistencies while working with it.
Q5 - It is easy to remember how to use it.
Q6 - It is easy to learn how to use the system.
Q7 - It does not work in the way I expect to.
Q8 - The information provided by the system is easy to understand.
Q9 - It gives me clear picture about the result of my previous action.
Q10 - At some point I was trying to find help or manual.
Q11 - I am able to complete my work quickly using the system.
Q12 - The system is pleasant to use

The realization of these changes will improve the implementation of System 1. Once all this changes have been adopted, we will develop System 2 by adding new functionalities to System 1 making System 2 able to support the contributions of big amounts of users when developing and validating tasks in different languages.

VI. NEXT STEPS

Our future research work is geared towards a purely crowd-based system. To get a substantial crowd engaged and to ensure the quality of the data, we plan to do research and development in two main research directions:

- **incentives**: to design and implement effective incentive mechanisms based on non-monetary motivation. We plan to harness human intrinsic incentives (such as enjoyment, recognition, autonomy, connectedness, meaningfulness, or the feeling of contributing towards the greater good). To design the appropriate incentives we need to address three important issues:

  a) Clear understanding and communication of desired behaviors. This means defining what exactly is expected as human contribution in terms of setting goals and actions required to achieve them. The extent to which humans perceive a task as a set of actions that lead to measurable result will influence their perception of being able to produce a meaningful outcome as their own.

  b) Understanding and aligning human incentives with these desired behaviors. Here we must acknowledge diversity in human motivation and its dependence on the context, for example in competence, enjoyment and connectedness. However, we will go a step further by analyzing incentives through human knowledge-building experience in the emerging field of buliding and maintaining lexical-semantic resources and trying to identify novel incentive dimensions.

  c) Obtaining desired behaviors through language development task design to achieve both effective task completion and human satisfaction. Instances of task design proved in practice, may be generalized as task templates to encourage shared models of incentives and improve computation quality.

- **Reputation**: to design and implement reputation schemes to properly evaluate the quality of users' contributions, thereby improving the quality of the resource. We measure the performance of a user in terms of number of words/senses/examples translated or validated in a specific time stamp. Later, the algorithm will be adapted to consider the type of task the user is performing or the quality of the result. In case of having more than one possible translation for the same lemma or sense or example we plan to use different techniques that have demonstrated to perform better than majority voting [18].

We will employ user studies to understand the effectiveness of the incentives and reputation mechanisms used in language development tasks and by observing the variation of user's performance across different languages and cultures we will change or adapt them.

The implementation of Systems 2 and 3 will be done progressively considering the gradual involvement of the crowd. We will build every system upon the improvement of the previous one and by the addition of new features to better manage crowd users and their contributions. Different tests will be run to assess the new functionalities. The analysis of the results of these tests will help us to improve the performance and appearance of the systems.

Conclusion

In this paper we have presented a scalable framework to build and maintain multilingual lexical-semantic resources. The framework is composed of three systems where users from the crowd are involved progressively.

We have presented a user study to validate the usability of the current implementation that will be improved after analyzing the results obtained. We will move to the next system by the addition of new functionalities that allow the management of crowds of users and their contributions.

Acknowledgments

The research described in this paper is part of the current research developed by the author as part of her PhD. This paper is a joint work with Fausto Giunchiglia, Gabor Bella and Mladjan Jovanovic to whom I express gratitude for collaboration and contribution. The work is supported by European Union's 7th Framework Programme project ESSENCE Marie Curie Initial Training Network (GA no. 607062).

References

[1] G. A. Miller, R. Beckwith, C. Fellbaum, and R. August, "Introduction to WordNet : An On-line Lexical Database," no. August, 1993.

[2] E. Pianta, L. Bentivogli, and C. Girardi, "MultiWordNet: developing an aligned multilingual database," *Proc. First Int. Conf. Glob. WordNet*, no. 1996, pp. 293–302, 2002.

[3] D. Tufis, "BalkaNet : Aims , Methods , Results and Perspectives . A General Overview," *Rom. J. Inf. Sci. Technol.*, vol. 7, pp. 9–43, 2004.

[4] L. E. and R. G. Gonzalez-Agirre A., "Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base," *Proc. Sixth Int. Glob. WordNet Conf.*, 2012.

[5] P. Vossen, "EuroWordNet General Document," pp. 1–108, 2002.

[6] N. Savage, "Gaining wisdom from crowds," *Commun. ACM*, vol. 55, no. 3, p. 13, 2012.

[7] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Commun. ACM*, vol. 54, no. 4, p. 86, 2011.

[8] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?," *Perspect. Psychol. Sci.*, vol. 6, no. 1, pp. 3–5, 2011.

[9] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk," *EMNLP '09 Proc. 2009 Conf. Empir. Methods Nat. Lang. Process.*, vol. 1, no. 1, pp. 286–295, 2009.

[10] J. McDonough Dolmaya, "Revision history: Translation trends in Wikipedia," *Transl. Stud.*, no. July, pp. 1–19, 2014.

[11] C. Biemann and V. Nygaard, "Crowdsourcing WordNet," *Proc. 5th Glob. WordNet Conf. Mumbai India ACL Data Code Repos. ADCR2010T005*, pp. 5659–5664, 2010.

[12] L. von Ahn and L. Dabbish, "Designing games with a purpose," *Commun. ACM*, vol. 51, no. 8, p. 57, 2008.

[13] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artif. Intell.*, vol. 193, pp. 217–250, 2012.

[14] D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli, "Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose," *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (ACL 2014)*, pp. 1294–1304, 2014.

[15] W. Bainbridge, "CAREER : Online Education as a Vehicle for Human Computation," pp. 1–2, 2015.

[16] I. Garcia, "Learning a Language for Free While Translating the Web. Does Duolingo Work?," *Int. J. English Linguist.*, vol. 3, no. 1, pp. 19–25, 2013.

[17] D. Bertram, "Likert Scales… are the meaning of life :," *Univ. Calagary, Dep. Comput. Sci.*, p. pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/, 2007.

[18] U. Endriss and R. Fernández, "Collective Annotation of Linguistic Resources: Basic Principles and a Formal Model," *Proc. 51st Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, pp. 539–549, 2013.