Building a Lexico-Semantic Resource Collaboratively

Mercedes Huertas-Migueláñez[1], Natascia Leonardi[2], Fausto Giunchiglia[1]

[1]University of Trento, [2]University of Macerata

E-mail: mdlm.huertas@unitn.it, natascia.leonardi@unimc.it, fausto@disi.unitn.it

Abstract

Multilingual lexico-semantic resources are used in different semantic services, such as meaning extraction or data integration and linking, which are essential for the development of real-world applications. However, their use is hampered by the lack of mantenance and quality control mechanisms over their content. The Universal Knowledge Core (UKC) is a multilingual lexico-semantic resource designed as a multi-layered ontology that has a language-independent semantic layer, the concept core, and a language-specific lexico-semantic layer, the natural language core. In this paper, we focus on expert-based, collaborative workflow for building and maintaining our resource through lexicalization and evaluation of language elements via a dedicated User Interface (UI). We have run a three-month study to analyze the feasibility of the proposed solution. We interviewed participants to obtain a comprehensive vision with respect to different aspects related to the way they interacted with the UI and how the content presented through it was perceived. We concluded that this collaborative experience fostered not only the implementation of a resource, but also an improvement of its functionalities, and, above all, it represented an example of effective knowledge sharing which opened up the way to a network of collaborative intelligence.

Keywords: multilingual resource, collaboration, knowledge sharing, user study

1 Introduction

Lexico-semantic resources, such as English WordNet (Miller 1995) and the corresponding parallel projects such as GlobalWordNet[2], are important to guarantee the presence of a language in our information society; for machine understanding related tasks, such as natural language processing (NLP) and machine translation (MT); and for people to learn and understand the lexico-semantic relations among language elements. However, the majority of these resources have some unresolved issues, such as content quality, i.e. typos or wrong translations (Zhang, Ojha & Giunchiglia 2017), or license restrictions, which can hamper their use and maintenance (Bond & Foster 2013).

Various people's contributions have been used to build and maintain linguistic resources. Wiktionary adopted crowdsourcing to build and maintain its content (Meyer & Gurevych 2012), and this allows the collection of data in a fast and cheap manner. However, the quality of the work produced by this method might be undermined by workers who are interested in the number of tasks completed rather than in the quality of the results (Eickhoff & de Vries 2013). Nevertheless, according to Morita and Ishida (2009), collaborative translation produces high-quality results. In order to successfully employ the metaphor of collaboration, we need to design systems that facilitate communication between people and organize them in teams with a range of expertise (Kittur et al. 2013). Furthermore, people should identify themselves with the group they collaborate with and believe that their effort is important for the community (Rashid et al. 2006; Munro 2010).

[1] The present study is the result of a close collaboration among the three authors. However, Mercedes Huertas-Migueláñez wrote Sections 1, 4 and 5. Natascia Leonardi wrote Section 3. Fausto Giunchiglia wrote Section 2. The Abstract and conclusions were a collaborative effort of the three authors.

[2] http://globalwordnet.org/ [last accessed 31-3-2018].

Whereas our long-term goal can be found elsewhere (Giunchiglia et al. 2015), in this paper we focus on the evaluation of the preliminary version of a tool to co-construct a high-quality multilingual lexico-semantic resource, the UKC (Giunchiglia, Batsuren & Bella 2017). Initially, we import freely available resources automatically. However, due to the complexity of the vocabularies, we involve experts to refine and maintain what we import. We selected the Italian language as our case study. The results of this preliminary study will be used to improve the current design of a dedicated UI and the collaboration pipeline.

The paper is organized as follows. Section 2 describes the UKC. In Section 3 we describe the Italian LKC. Section 4 presents the design of the UI. Section 5 reports on the study we conducted, and Section 6 concludes the paper.

2 The Universal Knowledge Core

The *Universal Knowledge Core* (UKC) is a knowledge base developed at the University of Trento. Just like in WordNet (Miller 1995), a vocabulary consists of *synsets*, *lemmas*, *word forms*, *senses*, and *examples*, which are representations of the sense in use. However, the UKC is different from WordNet, and the parallel projects, in that it features a language independent layer called the *concept core*. The *concept core* includes the lexico-semantic relations and provides mappings of common lexical elements from different languages, contained in the *language core*, to formal concepts (Giunchiglia, Batsuren & Freihat 2018). Every vocabulary is stored in a *Language Knowledge Core* (LKC). An LKC is a working copy of the UKC's concept core restricted to two vocabularies: English and another one. In our case study, we have chosen Italian. In Figure 1, we provide an example to illustrate how the UKC is organized. The English word *bike* has two meanings, as a verb and as a noun. They are represented by two single word synsets and are connected to the corresponding Italian words through their reference concepts. However, in Italian there is no lemma for the verb *to bike*, and therefore it will be represented as a *lexical gap,* which denotes missing lexicalization in a given language.

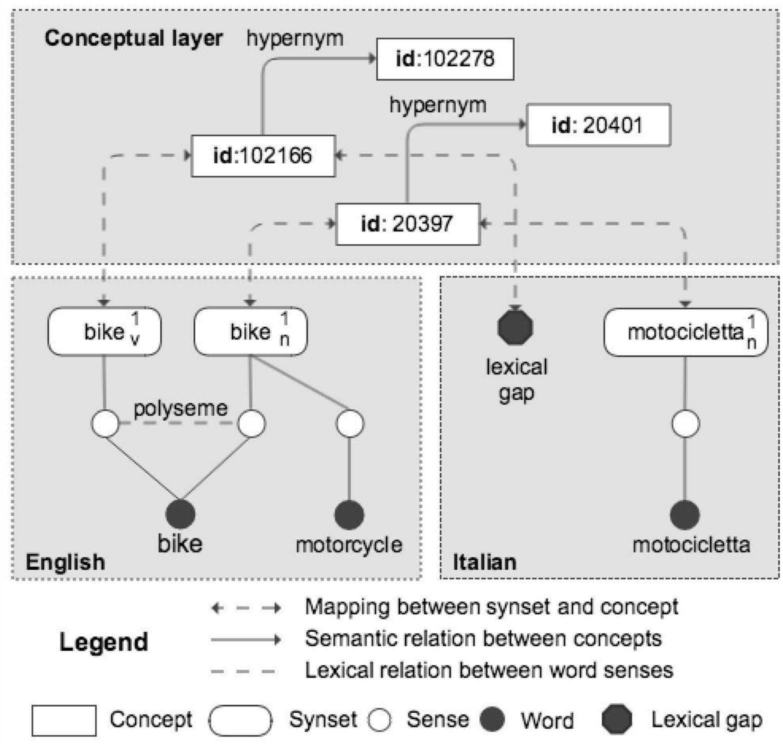


Figure 1: The UKC structure

The UKC is constantly growing by importing freely available resources that have been previously evaluated to keep our high-quality standards. Moreover, people are required to further populate the UKC, as we propose in this paper. Currently the UKC contains 335 languages, 1,333,869 words, 2,066,843 senses, and more than 120,000 concepts.

3 The Italian LKC

The development of the Italian LKC has been an experiment via collaboration between the University of Trento and the University of Macerata, which brought about the interdisciplinary merger of practices and methodological approaches of two distinct knowledge domains – namely Linguistics and Computer Science. A first experiment took place in the didactic context of a university seminar, where the students were the developers of the LKC and the professor was the instructor and final validator of their contributions. The local team was composed of five postgraduate students in Modern Languages for International Communication and Cooperation and an assistant professor in Computational Linguistics, whose roles were LKC developers and LKC validator/instructor, respectively. The students, whose mother-tongue was Italian, had a good knowledge of English; therefore, they correctly responded to the requirements of the UKC activity of equivalence compilation as required in their role. The professor, with an evaluator and trainer role, had a higher degree of expertise than the students regarding the background theory and procedures of the project, and was an expert in the fields of Linguistics, Terminology and Languages for Special Purposes (LSP), while the students could be considered semi-experts in these areas and in LSP Translation.

The inter-linguistic perspective and the analysis and definition techniques of the lexicon represented the common ground between the language developers' background knowledge and the requirements for the development of a lexico-semantic resource.

Linguistics and terminographic techniques (Wright & Budin 2001; Kockaert & Steurs 2015) were the principal skills required to accomplish the tasks oriented to the production of the bilingual lexico-semantic resource. Moreover, the students' expertise in terminology and terminography was applied both to the conceptual-semantic analysis of the lexicon and the elaboration of *intensional definitions* (Löckinger, Kockaert & Budin 2015). Indeed, the perspective adopted by the local team in their contribution to the LKC coincides with that used in the compilation of conceptually oriented lexical-semantic descriptions for the Italian equivalents of the English lemmas.

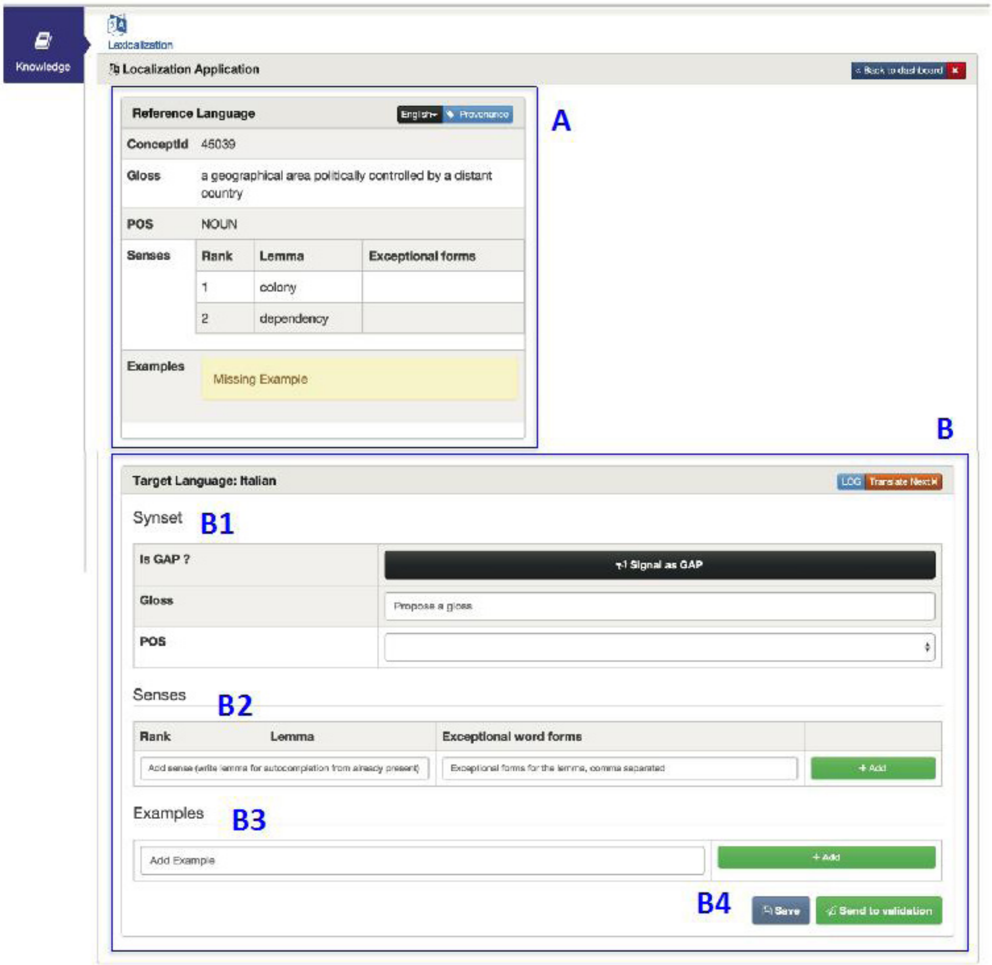
The shared knowledge between the two domains – i.e. terminology and LKC compilation – was used as a starting point for training in the current method, which entailed an adjustment of the developers' theoretical and practical approach to the activities of equivalence identification and definition writing. Therefore, the students' experience on (LSP) translation and terminological analysis turned into the ability to compile a computational lexico-semantic resource.

4 The User Interface

We present a preliminary version of the UI to facilitate contributions to build and maintain the UKC. In the UI presented in Figure 2, users can lexicalize English WordNet concepts into another language. In this preliminary approach, each contributing user will have a set of English WordNet concepts to lexicalize. The UI is divided into two parts:

- the top part contains the concept in the source language, Figure 2A, English in the example;
- the bottom part contains the corresponding empty fields to provide a lexicalization in the target language, Figure 2B.

If the concept to be lexicalized does not have a lexical equivalent in the target language, it can be marked by clicking on ‘Signal as GAP’, as in Figure 2B1. However, if there exists a lexicalization for the current concept, the user can complete the lexicalization by 1) adding the gloss and 2) selecting the corresponding POS from the pull-down menu, as in Figure 2B1; 3) adding a lexical equivalence for the lemma and an exceptional word form, that is irregular plural forms, irregular superlatives or irregular verb conjugations, when available as in Figure 2B2; and 4) adding an example as in Figure 2B3. When all the fields are completed, the user has to save the lexicalization first, ‘Save’ button, and then submit for evaluation, ‘Submit for validation’ button as seen in Figure 2B4. By clicking on the button ‘Translate Next’ a new English WordNet concept to be lexicalized will be available.



Reference Language

ConceptId	45039
Gloss	a geographical area politically controlled by a distant country
POS	NOUN

Senses

Rank	Lemma	Exceptional forms
1	colony	
2	dependency	

Examples: Missing Example

Target Language: Italian

Synset

Is GAP ?	
Gloss	Propose a gloss
POS	

Senses

Rank	Lemma	Exceptional word forms
Add sense (write lemma for autocompletion from already present)		Exceptional forms for the lemma, comma separated

Examples

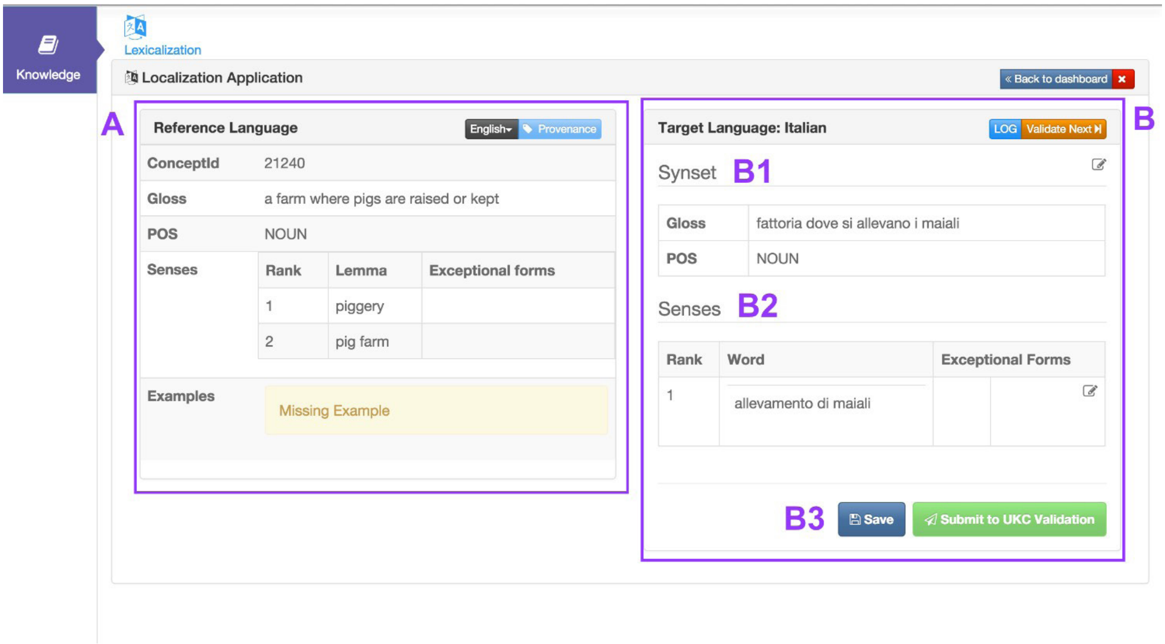
Add Example

Figure 2: UI design to complete a lexicalization. A corresponds to the source language part. B corresponds to the target language part. B1 corresponds to the gloss and POS of the word. B2 corresponds to the lexicalization of the word. B3 corresponds to the example of the word use. B4 buttons to save or submit the lexicalization.

In Figure 3 we present the UI to evaluate language elements. Again the screen is divided into two different areas:

- the left-hand side contains the concept in the source language, Figure 3A, English in this example;
- the right-hand side contains the lexicalized concept to evaluate, Figure 3B.

When a lexicalized concept is in the validation phase, it can be accepted by clicking on the ‘Submit for UKC validation’ or ‘Save’ if the concept needs further revision. By clicking the button ‘Validate Next’, a new concept to be evaluated will be shown. If a lexical element is marked as wrong the concept will be sent back to the lexicalization phase so that it can be revised.



Reference Language

ConceptId	21240
Gloss	a farm where pigs are raised or kept
POS	NOUN

Senses

Rank	Lemma	Exceptional forms
1	piggery	
2	pig farm	

Examples

Missing Example

Target Language: Italian

Synset

Gloss	fattoria dove si allevano i maiali
POS	NOUN

Senses

Rank	Word	Exceptional Forms
1	allevamento di maiali	

Figure 3: UI design to complete an evaluation. A corresponds to the source language part. B corresponds to the target language. B1 corresponds to the synset. B2 correspond to the senses and B3 are the buttons to save or submit the evaluation over a concept.

5 Study Design

We run a study on the Italian LKC to obtain a comprehensive vision with respect to 1) different aspects related to the evaluation of the UI; 2) how its content is perceived from the participants’ point of view, and; 3) the feasibility to build a lexico-semantic resource collaboratively. As far as we are concerned, only YARN (Braslavski, Ustaloc & Mukhin 2014) involved users in a pilot study. However, the methods used to capture participants’ opinions were not elaborated. In our study, we collected data using four methods to help us clarify contradictions in case any inconsistencies might be found. The approaches selected were: 1) think aloud (McDonald 2012) to understand and observe how they completed a translation task; 2) semi-structured interviews (Galletta 2013) to get a deeper insight on participants’ views and opinions on the UI and the content included in it; 3) desktop video-recording while interacting with the UI; and 4) a background questionnaire to obtain the demographics of the participants.

5.1 Evaluation

We granted access to the UI to the participants, and they decided how to organize the tasks they were asked to complete. We assigned them different sub-trees of the location domain related to region, geographical area, line, space, and point. Each of them contained between 75 and 127 nodes that corresponded to different concepts. After a period of three months, we met individually with the participants to interview them. All of them agreed to be voice-recorded and allowed the use of the resulting data for further analysis. The study was divided into three parts. Initially, we collected demographic information using a questionnaire. After that, we asked them to complete two translations using the UI while verbalizing what they were doing. Finally, we used semi-structured interviews to understand aspects related to the UI and how users perceived the content. Interviews were transcribed and thematically analyzed (Braun 2006).

5.2 Results

After three months, a total of 127 concepts were translated, among which 24 were classified as *lexical GAPs*. In spite of their other academic activities and the failures in the server where the UI was hosted, some of the students translated around half of the concepts they were assigned.

5.2.1 Interface Layout

With regard to the current implementation of the UI, the participants suggested that the design of the UI to lexicalize could be improved so that they always have at hand what needs to be lexicalized. Participant 1: "Sometimes I had to read the gloss several times, so I had to go up and down on the screen. I think it would be better to have everything on one screen". Their observation is corroborated after analyzing the video of their interactions, as they had to scroll up and down the page. Some of them pointed out that the visualization of the set of concepts they were assigned would have helped them to understand the relation among the words they had to lexicalize and, as a consequence, they could produce glosses accordingly. Participant 2: "Once I found the word 'colony' meaning colony of the United States and after, I found again 'colony' with a more general definition. Initially, I didn't know that I would find a second one, so I gave a more general definition in the first place. However, when I found it for the second time, I had to return and change the first definition." Some others felt that the way the tasks were presented was rather disorganized, making them feeling disoriented. Participant 4: "The fact that I could only see the current word instead of all the words I was assigned, made me feel that it was disorganized", Participant 6: "when I log in as a validator I get random entries".

5.2.2 Task Perception

The participants were enthusiastic about the tasks and the research process to find lexical equivalents, as it allowed them to learn nuances in the meanings of the words. Participant 4: "It helped to enrich my vocabulary and it is very useful to understand the language". In general, they felt that the experience of using a system like this was enriching and challenging. Participant 1: "it is very demanding because it needs a lot of research. It required a lot of time, but it was never boring. It was a very enriching task". As observed, the participants were very precise when completing their lexicalizations, as they were checking different monolingual dictionaries (LSP) corpora, in English and Italian, as well as trusted websites and images. They would only complete the translation when they really had a clear idea on how to add the lexical equivalent for the given concept.

5.2.3 Collaboration

The participants shared their experiences and doubts when their tasks were similar. Participant 4: "After finishing a task we compared what one has done with the others… contrasting always helps". When finding difficult concepts they asked the professor what steps to follow or what lexicalization for a specific object would be better. Participant 2: "with respect to conflictive cases, in order to create the gloss I would ask the professor what is a better option". The professor taught them how to produce a good lexicalization that would not be a literal translation of the English concept. She evaluated the lexicalizations produced, as well as replied to the different enquiries from the students so that there was a constant flow of information and feedback. Participant 6: “My students wondered whether they had to identify equivalents of the synsets whose definition might belong to another domain. I said ‘no’ because these concepts are not related to the domain of space/location”. Most of these communications were done face to face or via email.

6 Conclusions and Future Work

In this paper, we have presented a preliminary study to evaluate the design of a UI to maintain a multilingual lexico-semantic resource, the UKC, and whether collaboration among people is a feasible way to build and maintain a resource of these characteristics. We conducted a user study for three months in which six participants, five students and one professor of Linguistics, were involved. Although the total number of concepts lexicalized can be considered as low, mainly due to the failure of the system that forced the participants to access it in a discontinuous manner, this study helped us to obtain various improvements that could be introduced in the design of the UI, such as the inclusion of communication facilities and a of the redesign layout. We thus believe that it is possible to build a lexico-semantic resource based on collaboration. As shown here, the students were collaborating with each other while lexicalizing, as well as with the professor who was providing feedback and corrections to their work. This arrangement, where the professor is evaluating the semantic equivalences produced by the students, can be the seen as the most basic configuration. However, this can also be our baseline to understand if future collaboration settings, such as peer-to-peer, where students are lexicalizing and evaluating each other, could improve the results. In the future, we plan to import more freely available resources and involve contributors from different countries. We already have ongoing collaborations with groups in China, India, Mongolia, Romania, South Africa, and the United Kingdom (for Gaelic). The approach seems to be scaling without difficulty, at least from a technological point of view. The real difficulty is organizational: how to find and coordinate people from so many different countries working in parallel. The approach we are following is to build a community and a non-profit organization that will collaboratively manage the evolution of this resource.

References

Bond, F., Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *51st Annual Meeting of the Association for Computational Lingiustics. 4- 9 August 2013.* Sofia, Bulgaria.

Braslavski, P., Ustaloc, D., & Mukhin, M. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. In *Proceedings of the Demsontrations at the 14th Conference od the European Chapter of the Association for Computational Linguistics, ecacl2014, 26-30 April 2014.* Gothenburgh, Sweden.

Eickhoff, C., de Vries, A.P. (2013). Increasing cheat robustness of crowdsourcing tasks. In *Information Retrieval*, 16(2), pp 121-137.

Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication*. New York/London: NYU Press.

Giunchiglia, F., Jovanovic, M., Huertas-Migueláñez, M., & Batsuren, K. (2015). Crowdsourcing a large scale multilingual lexico-semantic resource. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP2015, 8-11 November 2015.* San Diego, USA.

Giunchiglia, F., Batsuren, K., & Bella, G. (2017). Understanding and Exploiting Language Diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI2017, 19-25 August 2017*. Melbourne, Australia.

Giunchiglia, F., Batsuren, K., & Freihat, A.A. (2018). One World – Seven Thousand Languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018.* Hanoi, Vietnam.

Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. & Horton, J. (2013). The Future of Crowd Work. In *Proceedings of the 2013 conference on Computer supported cooperative work, CSCW 2013, 23-27 February 2013*. San Antonio, TX, USA.

Kockaert, H.J., Steurs, F. (2015) (eds.). *Handbook of Terminology*. Vol. 1. Amsterdam/Philadelphia: John Benjamins.

Löckinger, G., Kockaert, H.J., & Budin, G. (2015). Intensional Definitions. In H.J. Kockaert, F. Steurs (2015), pp. 60-81.

McDonald, S., Edwards, H.M. & Zhao, T. (2012). Exploring Think-Alouds in Usability Testing: An International Survey. In *IEEE Transactions on Professional Communication*, 55(1), pp. 2-19

Meyer, C.M., Gurevych, I. (2013). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. Chapter 13 in S. Gragner, M. Paquot. (eds.) Electronic Lexicography. November 2012. Oxford: Oxford University Press, pp 259-251.

Miller, G. (1995). WordNet: a Lexical Database for English. In *Communications of the ACM*, 38(11), pp 39-41.

Morita, D., Ishida, T. (2009). Designing Protocols for Collaborative Translation. In *International Conference on Principles and Practice of Multi-Agent Systems, PRIMA2009, 14-16 Nov 2009.* Nagoya, Japan.

Munro, R. (2010). Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Translation Crowdsourcing for Translation. 31 October 2010, Denver, USA.*

Rashid, A.M., Ling, K., Tassone, R.D., Resnick, P., Kraut, R. & Riedl, J. (2006). Motivating Participation by Displaying the Value of Contribution. In *Proceedings of the SIGCHI conference on Human Factors in computing systems, CHI2006, 22-27 April 2006.* Montreal, Canada.

Wright, S.E., Budin, G. (2001). *Handbook of Terminology Management: Application-Oriented Terminology Management*. Vol. 2. Amsterdam/Philadelphia: John Benjamins.

Zhang, H., Ojha, S.R. & Giunchiglia, F. (2017). Finding errors in a Chinese lexico-semantic resource using GWAP. In *Proceedings of the IEEE Eleventh International Conference on Semantic Computing, ICSC2017, 30 January-1 February 2017.* San Diego, USA.