# Crowdsourcing a large scale multilingual lexico-semantic resource

**Fausto Giunchiglia, Mladjan Jovanovic, Mercedes Huertas-Migueláñez, Khuyagbaatar Batsuren**

Department of Information Engineering and Computer Science, University of Trento, Italy
{fausto.giunchiglia, mladan.jovanovic, mdlm.huertas, kbatsuren}@unitn.it

**Abstract**

Our goal is the construction, maintenance and evolution of a large-scale multilingual lexico-semantic resource, called *UKC* (for Universal Knowledge Core). Differently from previous approaches where similar resources were built by experts often on the basis of a corpus of reference documents, the UKC is constructed via *crowdsourcing*. We see language(s) as a live phenomenon that must be captured, studied and used, as much as possible in real time, while being generated and used by people.

## Introduction

The varying quality of existing lexico-semantic resources, such as WordNet (Miller 1995), influences the quality of services and applications that use them, such as meaning extraction, data integration and linking, or semantic search applications. BabelNet (Navigli & Ponzetto 2012) is a state of the art large multilingual semantic network created automatically by merging WordNet senses and Wikipedia entries. Expert annotators validated translations in five languages: Spanish, French, German, Italian and Catalan. Furthermore, *gamification* techniques (Vannella et al. 2014) have been applied to validate semantic relations and sense-image mappings.

The manual development of lexico-semantic resources is expensive in terms of required human power, but highly valuable with respect to the quality of the resource produced. *Crowdsourcing* has been extensively used in translation projects (Pavlick et al. 2014). The main motivation of our research relies on the study of how people use languages and how language evolves in time in different regions of the world. The expected outcome is a *scalable* lexico-semantic resource, not only in terms of size, e.g., number of words, but also in terms of the number and type of languages, including dialects; the evolution of the languages, including the most recent words; and diversity, including the less frequently used words, which may never make it to a dictionary.



The *Universal Knowledge Core* (UKC) (Tawfik et al. 2014) is a knowledge base developed at the University of Trento. The UKC presents a structure somewhat similar to that of WordNet and is designed as a multilayered ontology that has a language-independent semantic layer called the *concept core*, and a language-specific lexico-semantic layer where vocabularies are included. The UKC provides mappings of common lexical elements from different languages to formal concepts. The UKC is verticalized along language-specific vocabularies and involves, using the WordNet terminology, the generation of various language (related) elements: words, senses, synsets, , sets of synonymous words, glosses and examples of word usage (Miller 1995).

The verticalization process cannot 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 describes the main ideas underlying the implementation of this process.

## Crowdsourcing Language Development

We think of language development as the process of *translating* language elements from existing resources. By this we mean one of the following activities:

1. *Addition* of language elements. This activity can actually happen in two ways: *addition* of an element in a language as the lexicalization of a concept for which we have a word in another language (e.g., *pianta* in Italian and *Plant* in English); and *addition* of a lexical gap, namely of a word in a language for which there is no corresponding word in another language (e.g., there is no word in Italian to express the English word *biking*). Differently from WordNet, in the UKC lexical gaps are associated to a gloss;

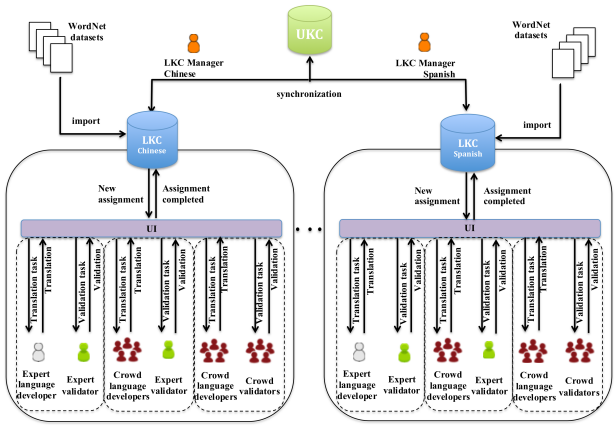
2. *Validation* of the correctness of the newly inserted language elements (e.g. the word spelling, the chosen example).

We implement these two operations as part of four pipelines which are activated in sequence, one for each language and which, once activated, can run in parallel:

1. vocabulary bootstrapping from existing WordNet-like resources, if allowed by their licenses;
2. expert-sourcing for the addition and validation of language elements;
3. crowd-sourcing for the addition of language elements and expert-sourcing for the validation;
4. crowd-sourcing for the addition and validation of language elements.

We define an *expert user* as a person *that we know that* her mother tongue is the language under development, that has a competent level of English, that has an extended knowledge of the two languages and whose activity is carried out in a controlled environment. We define a *crowd user* as a person whose level of expertise, in any of the competences described above, is unknown to us. A crowd user operates in an uncontrolled modality, we only require to login to be able to identify multiple sessions of the same person (useful for quality control but also user studies, both about single users and user groups).

The techniques above-mentioned are integrated in the framework below.




The framework implements a language development process for every Local Knowledge Core (LKC). An LKC is a working copy of the UKC's concept core restricted to two vocabularies: English and its correspondent localization. In this initial phase we use English as this is the most developed language. In the long term we plan to allow translation between any two languages. All registered users are assigned a role, where each role implements a part of the language development workflow as follows:

- *LKC developer*, who builds the target resource (e.g., by translating from a source or by providing lexicalizations from scratch);
- *LKC validator* who evaluates developers' work;
- *LKC manager* who is a trusted expert. There is one manager per language and will evaluate validators' work. The LKC manager is in charge of the overall process and, ultimately, is the person approving the synchronization and merging of the LKC with the UKC.

## First Results

So far we have imported 31 languages among which: Italian, Chinese, Spanish, Hindi, Portuguese and Arabic. The current system has more than 884.000 synsets and more than one million lemmas in total. We have also implemented a UI for experts to add and validate elements in their mother tongue. We have run a user study from December 2014 to February 2015 where a number of expert users from Italy, China, Mongolia and Bangladesh were asked to complete sets of tasks. In total they have developed around a hundred words, one hundred synsets and more than two hundred senses.

## Conclusions and Future Work

In this paper, we have presented a framework based on crowdsourcing to capture diversity in language into a multilingual lexico-semantic resource. The next step is to enable and activate the first step of the fourth, fully crowdsourced pipeline. Our short-term goal is to go live with the full system at the beginning of 2016. After that, our main aim will be to drive and coordinate the four development pipelines dealing with research issues such as: *extended use cases* which we will use to monitor the resource evolution, *incentives*, towards engaging users with the appropriate task, and *reputation*, for evaluating user performance.

## References

Miller, G. a., 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11), pp.39–41.

Navigli, R. & Ponzetto, S.P., 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193, pp.217–250.

Pavlick, E. et al., 2014. The Language Demographics of Amazon Mechanical Turk. Transactions of the Association for Computational Linguistics, 2, pp.79–92.

Tawfik, A., Giunchiglia, F. & Maltese, V., 2014. A Collaborative Platform for Multilingual Ontology Development., 8(12), pp. 3795–3804.

Vannella, D. et al., 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL *2014)*, pp.1294–1304.