Combining Gamification, Crowdsourcing and Semantics for Leveraging Linguistic Open Data

Antonio J. Roa-Valverde

STI Innsbruck,
Technikerstr. 21a, 6020 Innsbruck, Austria
antonio.roa@sti2.at

Abstract. In this paper we introduce Word Bucket, a mobile app that applies gamification to the problem of learning a foreign second language (FSL). Word Bucket consumes and produces linguistic data through user interaction, which can be used to improve available datasets relying on the power of the crowd. We describe the problems around handling "live" linguistic data and how semantic technologies can help to face the problem of data integration and its consumption in this specific scenario.

Keywords: linguistic data, data integration, language learning, gamification, mobile apps

1 Introduction

A recent survey conducted by the FP7 Project LIDER[1] shows that dictionaries, corpora and tokenizers are the most widely used linguistic resources by the community. Bilingual and multilingual dictionaries get a lot of traction among users. An example of this kind of resources is Wiktionary[2]. Wiktionary is an open source dictionary edited by the community. It offers data that is split into different language editions (one per supported language). For example, the English edition contains English descriptions of English words, but also of other languages. Terms existing in one edition can link to terms of different editions, creating a multilingual resource of available translations. The existence of abundant translations from a source language to a target language determines in part the quality and potential usage of these resources in practice. Unfortunately not all languages share the same level of support from the community, which leads to big quality differences among different Wiktionary editions.

In the recent years, due to the proliferation of the mobile platforms, many apps started to consume lexical data, including Wiktionary resources. Ranging from dictionaries to flashcards applications, most of these apps try to serve users as a tool for learning a foreign second language (FSL). Considering the impact of mobile apps among users, one could think about alternatives to improve the community support of lexical resources like Wiktionary. A possible way to increase

[1] https://www.w3.org/community/ld4lt/wiki/images/8/8e/Ld4lt-survey-apr14.pdf
[2] www.wiktionary.org

the user engagement is by combining FSL apps with gamification. Building on these ideas this paper presents:

- The use case of learning a FSL and how open linguistic data can be applied to this scenario.
- A discussion on the data heterogeneity issues around language resources from different providers and the efforts to find a standardized vocabulary for modeling and sharing linguistic data.
- The potential benefits of combining gamification and crowdsourcing for facilitating the task of learning a FSL, while at the same time producing data that can be reused for improving the quality of the original data sources.
- A data analysis approach based on ranking to ensure that users consume the data they expect.

2 Motivating Use Case

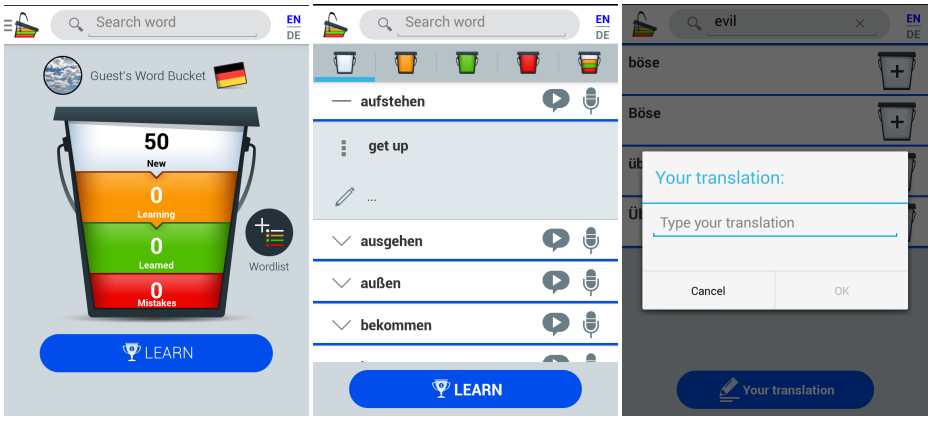
Learning a FSL is a task that implies perseverance and continuous motivation. From all the tasks involved in the process of learning a FSL, the acquisition of vocabulary is the one which is present in all stages. From beginners to advanced students, who command the structure of the language, the chances that new vocabulary is needed are high. This is not strange, since we are used to observe this fact in the development of our mother tongue too.

Dictionaries are without doubt the resources that best suit for vocabulary acquisition. While the amount of commercial dictionary publishers is endless, the proliferation in the last years of online collaborative projects like Wikipedia has originated similar approaches for the construction of dictionaries. Wiktionary can be seen as the leading lexical resource generated by the community.

Despite the Wiktionary content can be applied in many different areas of Natural Language Processing and Machine learning[3], there is still way to go in order to make the contribution of the Wiktionary community comparable to that of Wikipedia in terms of support and engagement. Even though it is clear that all the community would benefit from having better lexical resources, pushing the users to perform tasks like new content creation, translation or content curation is difficult as these tasks can result tedious and repetitive. A possible way of incentivizing users to accomplish this kind of tasks is by using gamification techniques. The basic idea of gamification is to hide the details of the real task, while generating an execution environment from which the user could directly benefit. Gamification pursues to incentivize the user to accomplish a task that otherwise she would not do. Following this line, we have developed Word Bucket, an app that applies gamification to the process of learning and reinforcing vocabulary in a FSL (Figure 1[4]).

[3] For an extended list of tasks where Wiktionary has been applied see "Wiktionary data in natural language processing" at http://en.wikipedia.org/wiki/Wiktionary

[4] Word Bucket also includes training functionality to help learning the stored vocabulary, however this part is not shown in the figure for brevity. For more details, we encourage the reader to download the app from `www.wordbucket.com`.



(a) Home (b) Explorer (c) Search

Fig. 1: Word Bucket's dictionary function

The relevance of Word Bucket relies on the way users interact with the data. On the first hand, Word Bucket is a consumer of lexical data, i.e. translations, by querying information from multilingual dictionary resources. Users can save associations of words in the target language they are learning, together with the respective translations in their mother language (Figure 1b). These associations are used to build different kind of games or tests that proof the knowledge of the learner. On the second hand, Word Bucket is a producer of data. Users can add their own translations to the app when there are not results for their query or these are just not appropriate (Figure 1c).

So far Word Bucket extracts only content from Wiktionary, which is still a resource under construction and presents data deficiencies. This lack is deeper for some languages than others. A possible way of enriching the original datasets is by integrating the user generated content back to the dictionaries. So far, in the current Word Bucket version, there is no possibility to reuse the user interaction for the benefit of the community. Afraid of this situation, we have started the development of a solution with the aim of feeding back the original dataset with data generated on the client side.

The public Wiktionary statistics[5] show that the engagement of the community has stabilized in the last 5 years. Figure 2 depicts the amount of active editors for the English Wiktionary along the time. There are not public statistics discerning between different kind of modifications, so we can not know the exact amount of changes that refer to translations only. Nevertheless, if we compare the graphic with the one in Figure 3, we can see that even the amount of daily Word Bucket users (for all the offered languages in Android) is still under the maximum number of English active editors, the potential amount of user

[5] http://stats.wikimedia.org/wiktionary/EN/Sitemap.htm

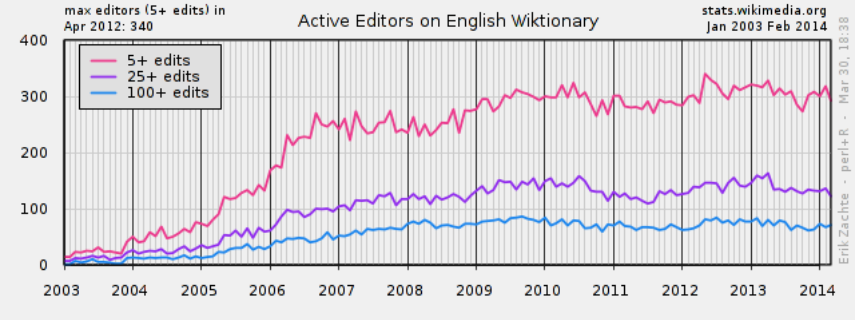


Fig. 2: Active editors on English Wiktionary

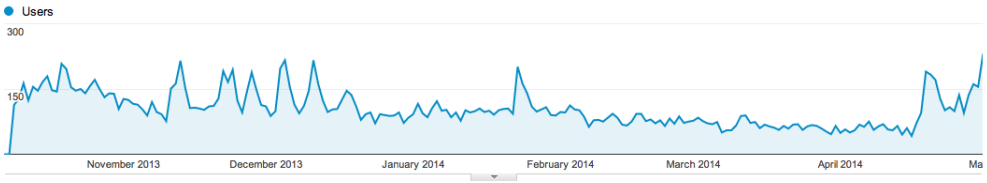


Fig. 3: Word Bucket daily active users (Android version) Oct 2013 - May 2014

generated translations could be of great benefit for the community. Moreover, the associations word-sense stored by users in their apps can be used as an indicator for creating scores in the translations. This can help the user to distinguish frequent translations from uncommon ones. At the time of writing this concept does not exist in Wiktionary, but it can be appreciated in other commercial resources like Google Translate[6].

We have identified the following requirements with the aim of improving the quality of the data and the user experience:

- Integrate other lexical resources to compensate the potential lack of quality. The data integration must be transparent to the client application.
- Use the lexical data generated on the client side to curate and enrich the original dataset.
- Model the user behavior and incorporate usage statistics that will improve the data consumption.

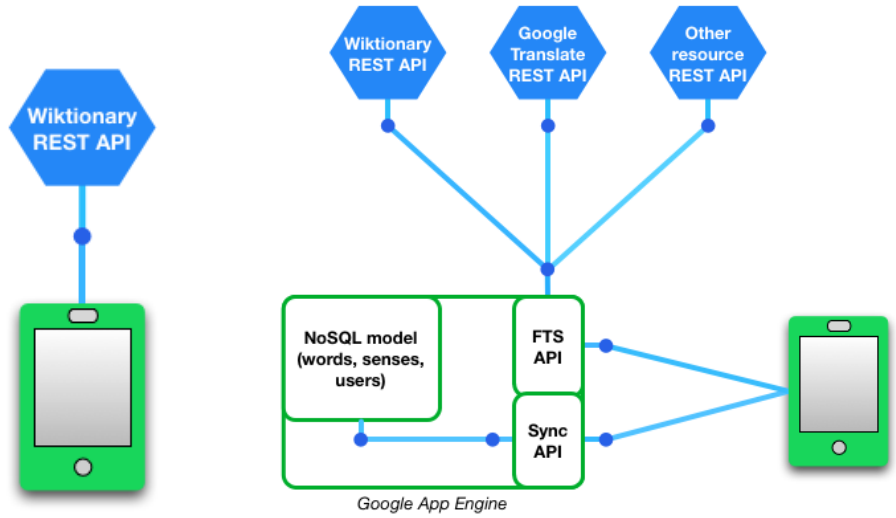
3 Design and Implementation

Along the different versions of Word Bucket we have tried to remove complexity from the device and implement richer functionality on the server side. In this way, we have modified how lexical data is consumed from one version to

[6] translate.google.com

another. In version 1.0 and version 2.0, we used a service federation approach. This means that data is consumed directly from the service providers by using their REST APIs. Figure 4a shows the original deployment, in which the device was responsible for implementing restful clients for each one of the integrated resources (in this version, Word Bucket only consumed data from Wiktionary).

As result of the REST requests, the services usually return JSON data, which needs to be parsed and converted to the internal data representation on the client side. The main drawback of this solution is that the data needs to be converted on the fly for every request. On the other hand, it allows us to have the latest lexical data offered by the resource publishers.



(a) Version 1.0 (b) Version 2.0

Fig. 4: Word Bucket service federation approach

An additional problem of the described deployment is the no possibility to share any of the user generated data. With the aim of fixing this problem, we developed a backend in version 2.0 as shown in Figure 4b. While still using a service federation, with this solution we were able to move all the data integration to the server side, removing complexity from the app package. The introduction of the backend allowed us to build our dictionary provider service, which certainly behaves like a proxy delegating client requests to the different resource providers. A great benefit of having this centralized proxy is the implementation of a full text search layer to homogenize the way we query the different lexical resources that we integrate or we might want to integrate in the future. In version 1.0, the

search strategy needed to be handled on the client side as part of the resource integration as well.

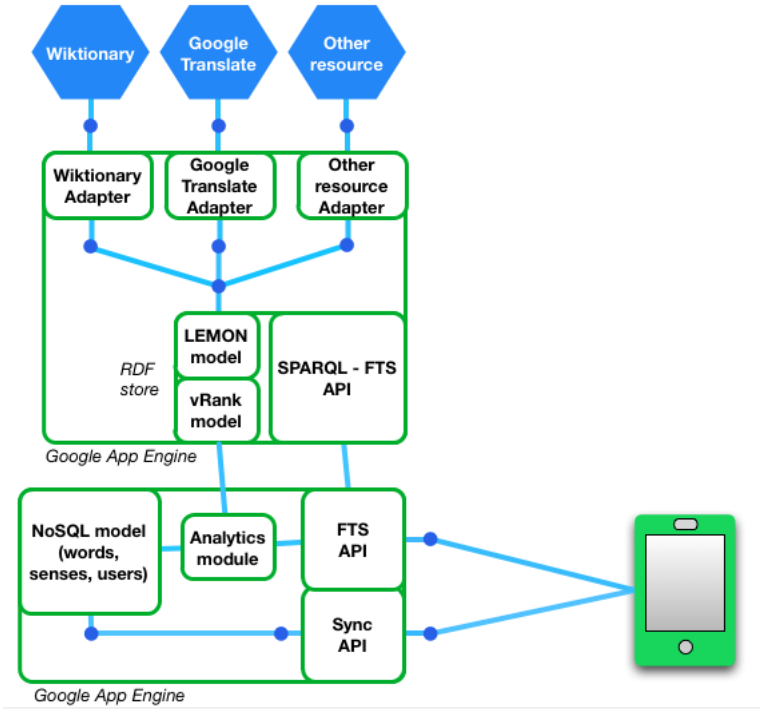


Fig. 5: Word Bucket data warehouse approach

With the introduction of the backend services in 2.0 we could start storing user data through a sync API, which allows the user to have a replica of their data on the cloud. As shown in 4b, all user generated data is kept independent of the original resources. As can be appreciated, with this approach the problem of data heterogeneity still remains open. In order to tackle this issue, we are currently working on the deployment of a new approach that builds on the use of semantic technologies (Figure 5). The main idea is to use a data warehouse solution in which the different lexical data is unified under a common format. As we already stated in section 2, when referring to the lexical data needs, Word Bucket focuses mainly on the consumption of translations in different languages. For this purpose, we will rely on the *lemon* model [2] together with the extension proposed by *DBnary* [1]. Following a similar strategy to the one described in [3] for the case of Wiktionary, we plan to build custom adapters for each resource. This task is precisely where the data integration will happen. The success of this task is crucial in order to expand the use of Word Bucket to other languages,

for which the current resources we are using do not show the expected level of quality. The solution we are preparing needs to be flexible enough to incorporate open and commercial resources interchangeably.

In order to incorporate user feedback to the lexical resources, we have built a module to collect usage statistics. The target of this module is to correlate the user queries with the best option from the list of possible translations. If the user decides to enter her own translation, this will also be considered for computing the statistics. Collecting this data from all the users will allow us to apply a ranking strategy and curate the dataset by removing noisy translations[7]. The rankings computed in this step will be made available as part of our data store. We will relay on the Vocabulary for Ranking (vRank) [5] for modeling the ranking information.

In the next subsections we describe further details on the implementation of the semantic backend solution.

3.1 Data consolidation

An issue that remains open in the field of computational linguistics is the development of knowledge artifacts and mechanisms to support the alignment of the different aspects of linguistic resources in order to guarantee semantic and conceptual interoperability in the linked open data cloud. Ontologies have proved to be of great use in achieving this goal and big efforts have been done in the Semantic Web community to address the conversion of datasets to RDF and its publication as linked data [6].

Recent initiatives like *lemon* [2] start to consolidate as *de facto* model to exchange linguistic data on the Web, which can be appreciated in the growing number of projects making use of it [1][3][4]. This fact is a first step towards solving the heterogeneity issues that exist when dealing with linguistics, specially lexical data coming from different providers.

As stated in [1], *lemon* is not sufficient for modeling bilingual dictionaries because it is not possible to represent translations. That is the reason why authors introduced *DBnary*[8] as a *lemon* extension. In order to avoid reinventing the wheel, we will make use of the *lemon* plus *DBnary* combination as part of our data model solution. Figure 6 shows the *lemon* model associated to the English term "bank". The complete list of senses associated to this term can be retrieved after executing the SPARQL query shown in Figure 8 towards the endpoint available at *http://kaiko.getalp.org/sparql*.

Figure 7 depicts an example of modeling a translation in *DBnary* and Figure 9 the respective SPARQL query to retrieve all Spanish translations associated to the term "bank".

[7] A similar strategy is already done by Google Translate.
[8] http://dbnary.forge.imag.fr/

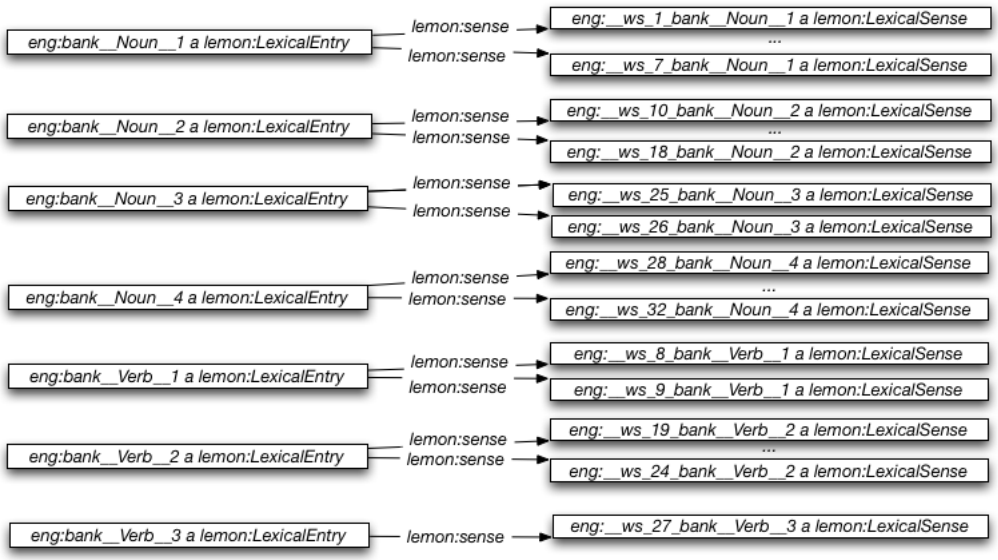


Fig. 6: Representation of the lexical term http://en.wiktionary.org/wiki/bank using *lemon*

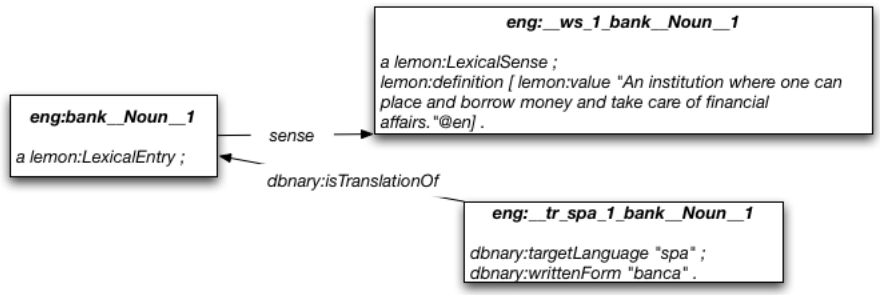


Fig. 7: *DBnary* representation of translations

3.2 Data analytics

An important part of our project focuses on incorporating user feedback to the data model. Applying similar mechanisms to those of Web Information Retrieval, we rank the available translations depending on the usage. As a rule of thumb, we consider that a translation has more chances to be right if many users keep it in their local Word Buckets. Following this principle, the scores associated to the translations will start converging in the datastore after heavy user interaction.

In case a translation does not exist, users have the possibility to create it according to their personal knowledge. These personal translations can be in-

```
select distinct ?word ?sense ?definition where {
        ?word a <http://www.lemon-model.net/lemon#LexicalEntry> .
        ?word <http://www.lemon-model.net/lemon#canonicalForm> _:a .
        _:a <http://www.lemon-model.net/lemon#writtenRep> "bank"@en .
        ?word <http://www.lemon-model.net/lemon#sense> ?sense .
        ?sense <http://www.lemon-model.net/lemon#definition> _:b .
        _:b <http://www.lemon-model.net/lemon#value> ?definition .
} order by ?word ?sense
```

Fig. 8: SPARQL query: get all senses for "bank"@en

```
select distinct ?w ?s ?t where {
        ?w a <http://www.lemon-model.net/lemon#LexicalEntry> .
        ?w <http://www.lemon-model.net/lemon#canonicalForm> _:a .
        _:a <http://www.lemon-model.net/lemon#writtenRep> ''bank''@en .
        ?s a <http://kaiko.getalp.org/dbnary#Translation> .
        ?s <http://kaiko.getalp.org/dbnary#isTranslationOf> ?w .
        ?s <http://kaiko.getalp.org/dbnary#targetLanguage>
                <http://lexvo.org/id/iso639-3/spa> .
        ?s <http://kaiko.getalp.org/dbnary#writtenForm> ?t .
} order by ?w ?s
```

Fig. 9: SPARQL query: get all Spanish translations for "bank"@en

corporated to the system and offered to the rest of users as part of the lexical dataset when they reach certain "credibility" threshold, i.e., many users have created the same translation in their buckets.

As stated previously, we need some kind of data model in order to make the translation scores persistent. For this purpose we have decided to use the Vocabulary for Ranking[9] (vRank) introduced in [5]. The aim of vRank is to provide data consumers with a standardized, formal, unambiguous, reusable and extensible way of representing ranking computations. Figure 10 shows an overview of vRank. *vrank:Rank* is an entity that formalizes the ranking scores associated to a data item. Anything that can be model in RDF can have an associated *vrank:Rank* instance. The flexibility of the model resides on relating different instances of *vrank:Rank* with a particular data item. A *vrank:Rank* by itself is meaningless. Therefore, *vrank:Rank* is related to *vrank:Algorithm*. In order to capture different executions we have added a timestamp to *vrank:Rank*. This property will allow us to monitorize how the translation scores evolve with the interaction. Figure 11 shows a complete example of our data model in turtle notation. Lines 45-51 show the the use of vRank.

4 Related Work

Previous works have been performed trying to improve the quality of linked data by using human contribution to achieve certain data related tasks. In [9], authors propose a framework based on Amazon's Mechanical Turk to achieve the execution of data related tasks by using the wisdom of the crowd. Related to the creation of linguistic data, authors in [10] apply crowdsourcing to address the creation of thesauri. In [14] authors proposed the MAPLE platform,

[9] http://purl.org/voc/vrank

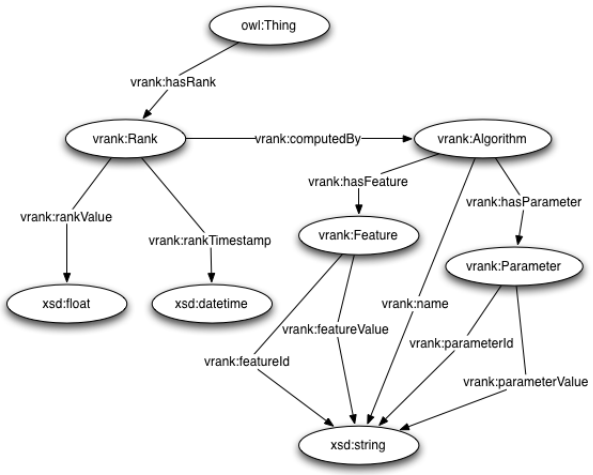


Fig. 10: vRank overview

which implements a Web adaptive learning solution based on RDF data models. MAPLE uses a reasoner to match tailored educational content with user profiles, in order to provide a custom learning experience. For this purpose, authors rely on the use of an extended RDF version of the LOM standard [15], that is used to describe the learning activities. All the user generated interaction is also modeled in RDF using an independent data model for later consideration during the matching phase. The different learning activities involve diverse multimedia content that is provided by an independent media delivery platform called NinSuna[10]. Authors state that the NinSuna platform is responsible for choosing the right media content according to the user's device platform, which aims to make adaptive mobile e-learning possible.

Closer to the idea of using games for generating data are the works described in [11], [12] and [13]. Specially in [16], von Ahn describes Duolingo, a mobile app based on gamification concepts that serves the users to learn a FSL while at the same time helps translating content publicly available on the Web. Unfortunately we could not find any references about the way Duolingo is handling the data and therefore we can not provide a comparison with Word Bucket. Other apps applying gamification to the problem of FSL learning are those provided by Busuu[11] and Babbel[12]. A first analysis of these apps reveals that the content they offer has been previously prepared and adapted for learning purposes, i.e., there is not direct consumption of any public data resource like in the case of Word Bucket.

[10] `http://ninsuna.elis.ugent.be`
[11] `www.busuu.com`
[12] `www.babbel.com`

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
@prefix lexvo: <http://lexvo.org/id/iso639-3/> .
@prefix dbnary: <http://kaiko.getalp.org/dbnary#> .
@prefix lemon: <http://www.lemon-model.net/lemon#> .
@prefix vrank: <http://purl.org/voc/vrank#> .
@prefix : <http://example.com/data#> .

:bank__Noun__1
       a       lemon:LexicalEntry ;
       dbnary:partOfSpeech "Noun" ;
       lexinfo:partOfSpeech
                lexinfo:noun ;
       lemon:canonicalForm
                [ lexinfo:pronunciation
                          "/ b k /"@en-fonipa ;
                  lemon:writtenRep "bank"@en
                ] ;
       lemon:language "en" ;
       lemon:sense :__ws_4_bank__Noun__1 ,
                                  :__ws_3_bank__Noun__1 ,
                                  :__ws_1_bank__Noun__1 ,
                                  :__ws_2_bank__Noun__1 .

:__ws_1_bank__Noun__1
           a       lemon:LexicalSense ;
           dbnary:senseNumber "1"^^<http://www.w3.org/2001/XMLSchema#int> ;
           lemon:definition
                    [ lemon:value "An institution where one
                                   can place and borrow money and take
                                   care of financial affairs."@en
                    ] .

:__tr_por_56_bank__Noun__1
            a        dbnary:Translation ;
            dbnary:isTranslationOf
                          :bank__Noun__1 ;
            dbnary:targetLanguage
                          lexvo:spa ;
            dbnary:writtenForm "banca" ;
            vrank:hasRank
                          :__rank_1__tr_spa_1_bank_Noun__1 .

:__rank_1__tr_spa_1_bank_Noun__1
    a vrank:Rank;
    vrank:hasRankTimestamp "2014-05-01T16:05:00"^^xsd:datetime;
    vrank:rankValue 0.83 .
```

Fig. 11: Data model example

5 Conclusions and Future Work

In this paper we have described the use case of learning vocabulary in a FSL by reusing lexical data from online providers. We have provided an overview of Word Bucket, a mobile app that combines gamification and crowdsourcing in this context, and discussed how the adoption of semantic technologies can help to solve the problem of lexical data heterogeneity and content enrichment.

Future work will focus on two main directions, namely, extending the offer of available languages and increasing the engagement of users. The first issue is an ongoing task since the beginning of the Word Bucket project. To increase the amount of languages supported, we need to incorporate specific resources targeting those languages. Due to the multilingual nature of Word Bucket, finding this kind of resources is not easy. The main problem resides in getting data containing bilingual translations, one for each pair of languages we would like to offer. A potential step towards a solution could be the addition of commercial dictionaries within Word Bucket. The integration of private and commercial

dictionary data into a global dataspace can open new business models for dictionary editors and service providers [7]. Far from the traditional offline model of "pay once and get it all", where printed dictionaries are the main purchased assets, the digital nature of online data opens new possibilities:

- Data licensing: users could be given access to certain parts of the data after purchasing a license token. Every content provider could establish its own terms, which could lead to the implementation of a marketplace strategy.
- Subscription model: users could get access to all available data for a short period of time after purchasing an access token.
- Pay per use: basically the same model followed by Google Translate, where users pay for a certain amount of consumed data. Different prices could be established according to the granted consumption quotas.

A requirement of this new approach is the need for providing provenance information within the data model, so that authorization mechanisms can be implemented. This problem has been already addressed by other authors in the context of pharmacological data [8].

Regarding the issue of engagement, a possible strategy would involve implementing new tests and minigames within the app. Engagement is directly related to the way users utilize the app. From a social perspective, mobile platforms have revolutionized how we interact with the information. Users carry their mobile devices most of the time. This factor facilitates the online presence of the user in comparison to using other devices like laptops or desktops. While mobile devices can be used everywhere, the second group of devices is only used in places like homes or offices. By using mobile apps, this fact can be properly exploited to generate content that otherwise could be hard for the user, not only because of the tedious of the task, but mostly because of finding the right time to accomplish it. Applied to the use case described in this work, lexical data can be enriched with user generated content (UGC) like notes, audio and images by using the right mechanisms to incentivize the user interaction. These annotations add an extra value to the original dataset and most importantly, they can be reused as part of the learning process.

Acknowledgments. The work described in this paper has been prepared in a close collaboration with *English Bubble Ltd.* Word Bucket and all the technical infrastructure around it is property of *English Bubble Ltd.* The authors acknowledge all the support provided by *English Bubble* team during the preparation of this paper. Word Bucket is the result of joining the effort and young spirit of very passionate people from different part of the globe. Special thanks go to Danny Smits and Robert Hanley.

We also thank Dieter Fensel and Miguel-Angel Sicilia for the support and feedback regarding this work.

References

1. Serasset, G. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. Semantic Web Journal, special issue on Multilingual Linked Open Data, 2014.
2. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web. 8th Extended Semantic Web Conference, 2011.
3. Hellmann, S., Brekle, J., Auer, S.: Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud. JIST 2012.
4. Westphal, P., Stadler, C., Pool, J.: Countering language attrition with PanLex and the Web of Data. Semantic Web Journal. 2012.
5. Roa-Valverde, A., Thalhammer, A., Toma, I., Sicilia, M.: Towards a formal model for sharing and reusing ranking computations. DBRank 2012.
6. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 1. Morgan & Claypool, 1st edition, 2011.
7. Cobden, M., Black, J., Gibbins, N., Carr, L., Shadbolt, N.: A research agenda for linked closed dataset. COLD 2011.
8. Carole A. Goble, Alasdair J. G. Gray, Lee Harland, Karen Karapetyan, Antonis Loizou, Ivan Mikhailov, Yrjn Rankka, Stefan Senger, Valery Tkachenko, Antony J. Williams, Egon L. Willighagen. Incorporating Commercial and Private Data into an Open Linked Data Platform for Drug Discovery. International Semantic Web Conference, 2013.
9. Simperl, E., Norton, B., Vrandecic, D.: Crowdsourcing Tasks within Linked Data Management. COLD 2011.
10. Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., Stuckenschmidt, H.: Crowdsourcing the Assembly of Concept Hierarchies. Joint Conference on Digital Libraries JCDL, 2010, Brisbane, Australia.
11. van Ahn, L., Dabbish, L. Designing games with a purpose. Communications of the ACM, 51(8):5867, 2008.
12. Siorpaes, K., Thaler, S., Simperl, E. SpotTheLink: A Game for Ontology Alignment. In Proc. 6th Conference for Professional Knowledge Management WM 2011, 2011.
13. Seneviratne, L., Izquierdo, E. An interactive framework for image annotation through gaming. International Conference on Multimedia Information Retrieval MIR 2010.
14. Van Deursen, D., Jacques, I., De Wannenmacker, S., Torrelle, S., Van Lacker, W., Montero Perez, M., Mannens, E., Van de Walle, R. A Mobile and Adaptive Language Learning Environment based on Linked Data. LILE 2011.
15. Nilsson, M., Palmer, M., Brase, J. The LOM RDF binding - principles and implementation. 3rd Annual Ariadne Conference (2003).
16. van Ahn, L. Duolingo: learn a language for free while helping to translate the Web. International conference on intelligent user interfaces, 2013.